Gradience in Grammar

Experimental and Computational Aspects of Degrees of Grammaticality

Frank Keller

PhD
University of Edinburgh
2000

Abstract

This thesis deals with gradience in grammar, i.e., with the fact that some linguistic structures are not fully acceptable or unacceptable, but receive gradient linguistic judgments. The importance of gradient data for linguistic theory has been recognized at least since Chomsky's *Logical Structure of Linguistic Theory*. However, systematic empirical studies of gradience are largely absent, and none of the major theoretical frameworks is designed to account for gradient data.

The present thesis addresses both questions. In the experimental part of the thesis (Chapters 3–5), we present a set of magnitude estimation experiments investigating gradience in grammar. The experiments deal with unaccusativity/unergativity, extraction, binding, word order, and gapping. They cover all major modules of syntactic theory, and draw on data from three languages (English, German, and Greek). In the theoretical part of thesis (Chapters 6 and 7), we use these experimental results to motivate a model of gradience in grammar. This model is a variant of Optimality Theory, and explains gradience in terms of the competition of ranked, violable linguistic constraints.

The experimental studies in this thesis deliver two main results. First, they demonstrate that an experimental investigation of gradient phenomena can advance linguistic theory by uncovering acceptability distinctions that have gone unnoticed in the theoretical literature. An experimental approach can also settle data disputes that result from the informal data collection techniques typically employed in theoretical linguistics, which are not well-suited to investigate the behavior of gradient linguistic data.

Second, we identify a set of general properties of gradient data that seem to be valid for a wide range of syntactic phenomena and across languages. (a) Linguistic constraints are ranked, in the sense that some constraint violations lead to a greater degree of unacceptability than others. (b) Constraint violations are cumulative, i.e., the degree of unacceptability of a structure increases with the number of constraints it violates. (c) Two constraint types can be distinguished experimentally: soft constraints lead to mild unacceptability when violated, while hard constraint violations trigger serious unacceptability. (d) The hard/soft distinction can be diagnosed by testing for effects from the linguistic context; context effects only occur for soft constraints; hard constraints are immune to contextual variation. (e) The soft/hard distinction is crosslinguistically stable.

In the theoretical part of the thesis, we develop a model of gradient grammaticality that borrows central concepts from Optimality Theory, a competition-based grammatical framework. We propose an extension, Linear Optimality Theory, motivated by our experimental results on constraint ranking and the cumulativity of violations. The core assumption of our

model is that the relative grammaticality of a structure is determined by the weighted sum of the violations it incurs. We show that the parameters of the model (the constraint weights), can be estimated using the least square method, a standard model fitting algorithm. Furthermore, we prove that standard Optimality Theory is a special case of Linear Optimality Theory.

To test the validity of Linear Optimality Theory, we use it to model data from the experimental part of the thesis, including data on extraction, gapping, and word order. For all data sets, a high model fit is obtained and it is demonstrated that the model's predictions generalize to unseen data. On a theoretical level, our modeling results show that certain properties of gradient data (the hard/soft distinction, context effects, and crosslinguistic effects) do not have to be stipulated, but follow from core assumptions of Linear Optimality Theory.

Acknowledgements

I am grateful to my supervisors Mark Steedman, Antonella Sorace, Matt Crocker, and Mats Rooth for continuous support and advice regarding the work reported in this thesis. Also, I would like to thank the members of my examination committee, Ellen Bard, Ewan Klein, Martin Pickering, and Hans Uszkoreit, for criticism and feedback on my research. The following people have provided valuable comments: Dora Alexopoulou, Ash Asudeh, Paul Boersma, Martin Corley, Alex Heneveld, Vasilios Karaiskos, Bob Ladd, Maria Lapata, Scott McDonald, Ineke Mennen, Paul Smolensky, Simone Teufel, and Jesse Tseng. Of course, the people on this list will not necessarily agree with all the claims made in this thesis, and the responsibility for any remaining errors is mine.

I have received important feedback when presenting my work at the universities of Potsdam, Saarbrücken, and Tübingen, and at the following conferences: Amlap-97, Amlap-99, CLS-98, Cogsci-00, and DGfS-99. Finally, I would like to thank the numerous subjects who participated in the experiments reported in this thesis. The financial support of DAAD, ESRC, and Studienstiftung is gratefully acknowledged.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Frank Keller*)

Contents

Chapter 1

Introduction

This chapter presents the motivation for studying gradience in grammar. It also summarizes the central claims put forward in this thesis and gives an overview of its structure.

1.1. Central Claims

Acceptability judgments are the basic data that linguists rely on to formulate their theories. In certain cases, these data fail to provide a clear-cut division between fully acceptable sentences and fully unacceptable sentences. Rather, relevant linguistic examples are *gradient*, i.e., they come in varying degrees of acceptability. The aim of the present thesis is to elucidate the status of these gradient examples and to show that a systematic, theoretically motivated treatment of gradience in grammar is possible.

This thesis puts forward four main claims. The first claim is that gradience is a systematic, pervasive grammatical phenomenon; gradient data occur in all parts of the grammar. We demonstrate this by conducting a series of experiments that cover all major syntactic modules and investigate representative syntactic phenomena in three languages. Our findings confirm that reliable gradient judgment data can be collected experimentally, and that such experimental data can yield insights that are not readily available from intuitive, informal linguistic judgments.

The second main claim is that all gradient phenomena share a common set of properties. These properties can be studied by investigating the effect that violations of grammatical constraints have on acceptability. We claim that constraint violations are ranked, i.e., they differ in seriousness. Also, constraint violations are cumulative, i.e., the degree of unacceptability increases with the number of violations. Furthermore, two types of constraints can be distinguished experimentally: soft and hard constraints. This dichotomy captures the intuition that certain linguistic constraints are binary, while others induce gradient acceptability judgments.

The third claim concerns the interplay between gradience and linguistic context. We

show that there is a systematic relationship between the soft/hard dichotomy and context effects. We provide support for the hypothesis that soft constraints are subject to contextual variation, whereas hard constraints are immune to context effects. This means that context effects can serve as a diagnostic for the soft/hard distinction.

The fourth central claim is that a model of gradient grammaticality can be devised that captures these experimental findings. We present a model that explains the empirical properties of gradient judgments (constraint ranking, cumulativity of violations, soft/hard dichotomy) and allows us to account for gradient data in a non-stipulative fashion by drawing on concepts from Optimality Theory.

1.2. Motivation for Investigating Gradience

This section discusses the four central claims of this thesis against the background of previous literature and motivates why these claims advance our understanding of gradience in grammar.

1.2.1. Theoretical Relevance of Gradient Data

The overarching assumption guiding this thesis is that gradient data can contribute to linguistic theory, over and above the binary judgments on which linguists traditionally rely. Often these judgments are not in fact binary, but constitute an idealization, i.e., the data are classified artificially into acceptable and unacceptable examples. In what follows, we argue that it is preferable to give up this idealization and develop a theory that permits us to analyze realistic, gradient data.

The potential benefits of a theory of gradient grammaticality include an expansion of the empirical base of linguistics and an increase of the predictive power of linguistic theory. As Hayes (1997b: 15) puts it: “Linguistics at present is *not hard enough*; we are not presenting our theories with sufficient demands to distinguish which ones are true. The task of analyzing data with gradient well-formedness puts a theory to a stiffer test.” Note that accounting for gradience was part of the research program of early generative grammar. Chomsky, for instance, insists that “an adequate linguistic theory will have to recognize degrees of grammaticalness” (Chomsky 1975: 131), based on the observation that “there is little doubt that speakers can fairly consistently order new utterances, never previously heard, with respect to their degree of ‘belongingness’ to the language” (Chomsky 1975: 132).

The present thesis explores these conjectures by demonstrating that the use of gradient data can indeed contribute new insights to linguistic theory. We investigate a wide variety of linguistic phenomena, taken from all major syntactic modules and from several languages. Our experimental results show that by taking gradient judgment data into account, we can both discover new linguistic facts that have eluded the conventional, informal approach to data collection, and resolve data disputes that exist for certain linguistic phenomena in the literature.

The underlying hypothesis is that such disputes arise because conventional linguistic analysis fails to do justice to the gradient nature of these phenomena, both in its data collection methodology and in its analytic approach.

Note that there is an important methodological caveat here. Arguably, the aim of formulating precise, testable theories of linguistic competence is at the heart of the generative enterprise. We have to make sure that this aim carries over to an extended theoretical framework that is capable of dealing with gradience. In other words, we have to make sure that a *formal* theory of gradience is possible, countering "[c]ritics of generative grammar [who] might take the existence of gradient well-formedness judgments as an indication that the entire enterprise is misconceived [...]. In this eliminativist view, gradient well-formedness judgments constitute evidence that generative linguistics must be replaced by something very different, something much 'fuzzier' " (Hayes 2000: 88). We follow Hayes (1997b: 15) in adopting the guiding assumption that "we don't have to trash existing theories of what constraints are like just to get gradient well-formedness". The present thesis aims to develop a grammatical framework that is permissive enough to account for gradient data without idealizing it, but restrictive enough to allow us to formulate precise, testable linguistic analyses. We show that such a framework can be designed as an extension of an existing theoretical approach, viz., Optimality Theory.

1.2.2. Empirical Properties of Gradient Data

Apart from advancing the understanding of gradience in grammar, the present thesis also makes a contribution to linguistic methodology. This contribution is a conservative one, which means that we accept the established approach of relying on native speakers' judgments as primary evidence for linguistic theory. We simply extend this approach from binary to gradient judgment data. This extension requires the use of experimental methods to obtain judgments, as informal procedures are not reliable for gradient data (see Sections 2.3 and 2.4). Also, the move from binary to gradient data makes it necessary to refine existing theoretical concepts (specifically the notion of constraint ranking, see Chapter 6).

Linguistic judgments are a fairly well studied behavioral phenomenon, and a wide range of psychological factors have been identified as having an influence on the judgment process, including task-related factors such as measurement scale, instructions, order of presentation, and subject-related factors such as field dependence, handedness, and literacy (see the review in Section 2.3). The present thesis, however, is not concerned with the effect of such *extra-linguistic* factors, but investigates how *linguistic* factors influence the degree of acceptability of a linguistic structure.

Two types of linguistic factors can be distinguished: performance factors, which are involved in language processing (including parsing, generation, and acquisition), and competence factors, which characterize the knowledge of language of a speaker (see Section 2.2.2 for a more extensive discussion). The present thesis will focus on competence factors, i.e., factors

which belong to the domain of linguistic theory and can be couched in terms of constraints, principles, or rules as they are typically postulated by linguists.

Note that previous experimental studies on competence effects in gradience are largely absent (with the exception of some early studies, see Chapman 1974; Coleman 1965; Marks 1967; Stolz 1967, which were mainly concerned with the influence of the type and number of rule violations on acceptability). The present thesis attempts to fill this gap by providing a systematic experimental investigation of how competence factors interact to determine the degree of acceptability of a linguistic structure (see Chapters 3 and 4). At the same time, we try to keep the influence of extra-linguistic factors as constant as possible by using standard techniques for the design and evaluation of psycholinguistic experiments (see Section 2.4).

It is important to emphasize that the research reported in this thesis is not psycholinguistic in nature; we are not concerned with linguistic performance, i.e., we make no claims about human language processing. We avail ourselves of experimental tools standardly used in psycholinguistics, but our focus is on linguistic theory, and the central tenet of our investigation is to extend the empirical scope and the theoretical reach of models of linguistic competence.

1.2.3. Gradient Data and Context

When linguists are confronted with a gradient datum, e.g., with a sentence *S* that is of reduced acceptability, but not fully unacceptable, they often resort to an argumentation like the following. They try to find a specific context *C* in which *S* is fully acceptable (or at least of increased acceptability). Having found such a context, they conclude that the structure instantiated by *S* is grammatical, a fact from which certain theoretical conclusions can be drawn. This strategy implies that "*S* is acceptable" actually means "there is a context in which *S* is acceptable". However, such an approach fails to recognize the distinction between sentences that are acceptable *without* requiring a specific context, and ones that are *only* acceptable in a specific context. This situation is recognized by Chomsky (1964), who states rather polemically:

> Linguists, when presented with examples of semi-grammatical, deviant utterances, often respond by contriving possible interpretations in constructed contexts, concluding that the examples do not illustrate departure from grammatical regularities. This line of argumentation completely misses the point. It blurs an important distinction between a class of utterances that need no analogic or imposed interpretation and others that can receive an interpretation by virtue of their relations to properly selected members of this class.
>
> (Chomsky 1964: 385)

There is, however, a research tradition that diverges from mainstream linguistics in that it recognizes the theoretical importance of the distinction between sentences that are acceptable per se and sentences that are acceptable only in a certain context. This line of research, initiated by

Lenerz (1977) and Höhle (1982) (among others) and later taken up by Müller (1999), concerns itself with the influence of context on word order. This influence is captured by the notion of *markedness*, which Höhle (1982: 102, 122) defines as follows: a sentence S_1 is less marked than a sentence S_2 if it can occur in more context types than S_2.

It has been proposed that this notion of markedness corresponds to speakers' intuitions about gradient acceptability, i.e., that "relative degrees of markedness can be empirically determined [...] either by directly invoking speakers' judgments, or by adhering to the number of context types in which the candidate [i.e., the sentence] is possible" (Müller 1999: 782f). Note however, that this approach remains speculative, Müller (1999) does not refer to experimental evidence to show that the number of contexts in which a sentence can occur correlates with its relative acceptability. Furthermore, Müller's (1999) approach can be regarded as circular: "number of contexts" is an intuitive notion that has to be judged by native speakers; hence he only defines one type of intuitive judgments in terms of another one.

The present thesis adopts an operational definition of the interaction of context and acceptability to address both Chomsky's (1964) criticism of current linguistic practice and the shortcomings of the markedness approach to gradience. This is achieved by introducing the notion of *context dependence* of linguistic constraints (see Chapter 4): a constraint is context-dependent if the degree of unacceptability triggered by its violation varies from context to context. The context dependence of a constraint can be determined experimentally and receives a precise interpretation in the model of gradience put forward in Chapter 6. Furthermore, we provide evidence for the hypothesis that only certain linguistic constraints—soft constraints—are context-dependent.

The research in this thesis is only concerned with the *linguistic* context of an utterance. By linguistic context we mean the linguistic material that precedes the sentence under investigation. We do not deal with effects from the extra-linguistic context, which are well-attested in the experimental literature on linguistic judgments (see Section 2.4 for an overview).

1.2.4. Modeling Gradient Data

Apart from contributing to the understanding of the experimental properties of gradient linguistic data, the present thesis is also concerned with how these data can be accounted for in a theoretically motivated way, i.e., how existing grammar models can be extend to accommodate gradience.

The interest in modeling gradience in grammar goes back to early generative linguistics, where the relevance of gradient data was recognized, and several attempts at modeling it were made (Chomsky 1955, 1964, 1965; Katz 1964). Bolinger (1961a), who also introduced the terms "gradience" and "gradient", provided a wealth of evidence showing that linguistic notions can be continuous, rather than discrete. His argumentation was mainly aimed at phonology, but Bolinger (1961b) later extended it to syntax. A similar line of research was

pursued by a group of generative grammarians working in the framework of Fuzzy Grammar, which is based on the assumption that linguistic categories are not discrete, but are organized hierarchically and annotated with application probabilities (Lakoff 1973; Mohan 1977; Quirk 1965; Ross 1972, 1973a,b). A related approach was proposed in sociolinguistics in the form of the Variable Rules framework developed by Cedergren and Sankoff (1974) and Labov (1969).

After the early surge of interest in gradience in grammar in the generative tradition, a remarkable gap in the literature occurred from the mid 1970s until the mid 1990s. No significant research on models of gradience seems to have taken place in this period. Arguably, this was a consequence of the failure of early attempts at modeling gradience to yield insightful theoretical results that simulated further research. It can be assumed that the reason for this lack of progress was the absence of adequate empirical and theoretical tools for the systematic investigation of gradience.

Two innovations rekindled the interest in gradience in the mid 1990s. One was the advent of Optimality Theory (Prince and Smolensky 1993, 1997) as a new theoretical framework. The other one was the availability of magnitude estimation as a new way of collecting judgment data (Bard, Robertson, and Sorace 1996; Cowart 1997). Optimality Theory assumes that constraints are inherently ranked and violable, it is based on an intrinsically *relative* notion of grammaticality, and therefore provides the conceptual repertoire for tackling the issue of gradience in a principled way. Magnitude estimation, on the other hand, affords linguists a tool for measuring judgments in a fine-grained and fully reliable way; it makes it possible to obtain data that goes beyond traditional informal data collection and puts the study of gradience on a sound experimental footing.

These new tools for the study of gradience triggered a surge in interest in the theoretical and empirical aspects of gradient data. The most important contributions were made by Boersma, Hayes, and colleagues (Boersma 1998, 2000; Boersma and Hayes 2001; Hayes 1997b, 2000; Hayes and MacEachern 1998), Cowart and colleagues (Cowart 1989a, 1994, 1997; Cowart, Smith-Petersen, and Fowler 1998; McDaniel and Cowart 1999), Müller (1999), Gordon and Hendrick (1997, 1998a,b,c), and Sorace and colleagues (Bard et al. 1996; Sorace 1992, 1993a,b, 2000; Sorace and Cennamo 2000; Sorace and Vonk 1998).

This thesis is part of this new generation of studies that tackle gradience using innovative experimental and conceptual tools. It relies extensively on magnitude estimation as a means of collecting reliable gradient judgment data, and uses Optimality Theory as a basis for formulating a model of gradience that is both grounded in linguistic theory and backed up by extensive experimental studies.

1.3. Overview of the Thesis

This thesis is divided into four parts: a background part (Chapter 2), an experimental part (Chapters 3 and 4), a methodological part (Chapter 5), and a theoretical part (Chapters 6 and 7).

Chapter 2 spells out the background assumptions on which this thesis rests. In particular, it provides an overview of the problems connected with linguistic judgments in general, and with gradient judgments in particular. The chapter also discusses the competence/performance distinction and how it applies to gradient phenomena. Finally, an overview of Optimality Theory and an introduction to the magnitude estimation paradigm is provided.

Chapter 3 presents a series of experiments that aim to establish a number of general properties of gradient linguistic judgments. These experiments deal with unaccusativity/unergativity, extraction, binding, and word order, and the aim is to investigate how constraint ranking, constraint type, and constraint interaction determine the degree of grammaticality of a given linguistic structure. The experimental findings indicate that there are two types of constraint violations: soft constraint violations that cause only mild unacceptability and induce gradience, and hard constraint violations that lead to strong unacceptability and are manifested in binary acceptability judgments. For both types of constraints, violations are cumulative, i.e., the unacceptability of a structure increases with the number of constraints it violates. The soft/hard dichotomy then motivates the study of the interaction of gradience and context in Chapter 4.

Chapter 4 reports a series of experiments on gapping, extraction, and word order that confirm the basic observations that constraints are ranked and that constraint violations are cumulative, but also provide additional evidence for the hard/soft dichotomy. The chapter presents crosslinguistic data on word order that makes it possible to investigate the crosslinguistic behavior of hard and soft constraints. Furthermore, it is shown that context effects are a powerful diagnostic of constraint type and results are presented that indicate that soft constraints are subject to context effects, while hard constraints are immune to contextual variation.

Chapter 5 discusses the problems and opportunities that arise from web-based experimentation, the methodology used for the experimental studies presented in Chapters 3 and 4. The chapter explains the software that was used for these experiments and the safeguards that were put in place to ensure the authenticity of the data obtained over the web. A number of experiments are presented which demonstrate the reliability and validity of web-based studies. This includes the web-based replication of the results of a lab-based study and a questionnaire-based study.

Chapter 6 develops Linear Optimality Theory, a model of gradient grammaticality that borrows central concepts from Optimality Theory, a competition-based grammatical framework. Linear Optimality Theory is motivated by the experimental results on constraint ranking and the cumulativity of violations, as demonstrated in Chapters 3 and 4. The core assumption

of Linear Optimality Theory is that the relative grammaticality of a structure is determined by the weighted sum of the violations it incurs. It is demonstrated that the parameters of the model (the constraint weights), can be estimated using Least Square Estimation, a standard model fitting algorithm. It is also shown that Standard Optimality Theory is a special case of Linear Optimality Theory.

Chapter 7 shows the validity of Linear Optimality by presenting two types of modeling studies: three small scale proof of concept studies that illustrate how specific properties of gradient data are accounted for by Linear Optimality Theory, and two larger, more realistic modeling studies that illustrate the interaction of a number of properties of gradient data. Throughout this chapter, Least Square Estimation is employed to determine model parameters (i.e., constraint ranks) from experimentally collected data. Crossvalidation is used to demonstrate that the predictions of a model generalize to unseen data. On a theoretical level, the modeling results show that certain properties of gradient data (the hard/soft distinction, context effects, and crosslinguistic effects) do not have to be stipulated, but follow from the core assumptions of Linear Optimality Theory.

1.4. Collaborations and Published Work

Three of the experiments reported in this thesis are the result of collaborations: Experiment 5 was conducted in collaboration with Ash Asudeh (Stanford); Experiments 11 and 12 were conducted in collaboration with Theodora Alexopoulou (Edinburgh).

Some of the material presented in this thesis has been published or is presently under review for publication; this applies to Chapter 2 (Keller 1999), Chapter 3 (Keller and Sorace 2000), Chapter 4 (Keller 2000, 2001; Keller and Alexopoulou 2001), and Chapter 6 (Keller 1998; Keller and Asudeh 2000).

Chapter 2

Background

The present chapter spells out the background assumptions on which this thesis rests. We provide an overview of the methodological issues connected with linguistic judgments in general, and with gradient judgments in particular. We also discuss the competence/performance distinction and how it applies to gradient phenomena. Finally, we introduce the magnitude estimation paradigm and give an overview of Optimality Theory.

2.1. Introduction

The data on which linguists base their theories typically consist of acceptability judgments, i.e., of intuitive judgments of the well-formedness of utterances in a given language. When a linguist obtains an acceptability judgment, he or she performs a small experiment on a native speaker; the resulting data are behavioral data in the same way as other measurements of linguistic performance (e.g., the reaction time data used in psycholinguistics). However, in contrast to experimental psychologists, linguists are generally not concerned with methodological issues, and typically none of the standard experimental controls are imposed in collecting data for linguistic theory.

Recently, there has been growing interest in the psychological aspects of linguistic judgments. A number of researchers have set out to investigate the experimental properties of acceptability judgments, as well as the implications that such experimental findings might have for linguistic methodology. The present chapter tries to provide an overview of the most relevant results, drawing mainly on the monographs by Schütze (1996) and Cowart (1997), as well as on the seminal article by Bard et al. (1996). Schütze (1996) aims to show that the methodological negligence that characterizes the bulk of linguistic research can seriously compromise the data obtained, and argues for a more reliable mode of experimentation, similar to the one standardly used in experimental psychology. The contributions by Bard et al. (1996) and Cowart (1997) are complementary to Schütze's (1996) more theoretically oriented discus-

sion. Both studies propose new procedures for eliciting acceptability judgments by drawing on methods from experimental psychology, and show how reliable, delicate data can be obtained using these procedures.

This chapter is organized as follows. Section 2.2 analyzes the practice of using acceptability judgments as evidence in linguistics. It also discusses the competence/performance dichotomy and its application to gradient judgments. Section 2.3 surveys the non-linguistic factors that can influence acceptability judgments, with special emphasis on the role of measurement scales and instructions. Based on this discussion, Section 2.4 discusses Schütze's (1996) recommendations for eliciting reliable judgment data, and explains how these recommendations are implemented in the present thesis. Section 2.5 elaborates on this by providing a description of the magnitude estimation paradigm used in Chapters 3–5. Finally, Section 2.6 gives an overview of Optimality Theory, the theoretical framework that this thesis builds on.

2.2. Acceptability Judgments and Linguistic Theory

This section deals with the role of acceptability judgments in linguistic theory and argues that in order to obtain reliable data, we have to pay attention to the psychological properties of acceptability judgments. We also explore the competence/performance dichotomy that underlies most of the work in generative linguistics, and discuss its application to gradient linguistic judgments.

2.2.1. Judgments as Evidence for Linguistic Theory

Acceptability judgments by native speakers are generally accepted as the main type of evidence for linguistic theory. The use of judgment data is typically justified by a set of key arguments that Schütze (1996: 2) summarizes as follows:

- Acceptability judgments allow us to examine sentences that rarely occur in spontaneous speech or corpora.
- Judgments constitute a way of obtaining negative evidence, which is rare in normal language use.
- In observing naturally occurring speech data, it is difficult to distinguish errors (slips of the tongue, unfinished utterances, etc.) from grammatical production.
- The use of acceptability judgments allows us to minimize the influence of communicative and representational functions of language. Judgment data allow us to study the structural properties of language in isolation.

This set of advantages explains the popularity of acceptability judgments as primary data for linguistic theory. However, as Schütze (1996) argues, judgment data are often used by linguists

in a dangerously uncritical fashion. “In the vast majority of cases in linguistics, there is not the slightest attempt to impose any of the standard experimental control techniques, such as random sampling of subjects and stimulus materials or counterbalancing for order effects” (Schütze 1996: 4). Linguists typically rely on a naive, intuitive way of collecting judgments, ignoring psycholinguistic findings that show that acceptability judgments are subject to a considerable number of biases, for which a naive methodology fails to control (see Section 2.3 for details). “In the absence of anything approaching a rigorous methodology, we must seriously question whether the data gathered in this way are at all meaningful or useful to the linguistic enterprise” (Schütze 1996: 5).

Schütze (1996) also points out that current linguistic research makes crucial use of subtle (and thus potentially controversial) judgments; it does not confine itself to cases of clear acceptability or unacceptability (which arguably can be established without using an experimental methodology): “The days are over when linguistics had more than enough to worry about with uncontroversial, commonplace judgment data, and the sophisticated and complex judgments now in use by theoreticians assume much about human abilities that remains unproven, even unscrutinized” (Schütze 1996: 9). To substantiate this claim Schütze (1996: 36–38) discusses the use of subtle judgments in the widely cited studies by Aoun, Hornstein, Lightfoot, and Weinberg (1987), Belletti and Rizzi (1988), and Lasnik and Saito (1984).

Belletti and Rizzi’s (1988) study is particularly interesting as it makes extensive use of gradient acceptability judgments, de facto employing a seven point scale for acceptability. However, the authors fail to provide an explicit account of degrees of grammaticality:

> But there is no general theory of which principles *should* cause worse violations. The theory makes no prediction about the relative badness of, say, θ-Criterion versus Case Filter violations, let alone about how bad each one is in some absolute sense. The notion of relative and absolute badness of particular violations is ad hoc, and is used in just those cases where it is convenient.
>
> (Schütze 1996: 43)

This problem is not limited to Belletti and Rizzi’s (1988) paper. Even a well-known syntax textbook such as Haegeman’s (1994) suffers from similar difficulties. Haegeman (1994) makes extensive use of intermediate acceptability ratings, which in the absence of clear criteria on how to record and interpret intermediate judgments can lead to serious inconsistencies, as Bard et al. (1996) discuss in some detail.

These examples indicate that the use of gradient acceptability judgments is common in the linguistic literature. However, the reliance on subtle data is not matched by the necessary concern for experimental methodology. Also, the theoretical treatment of gradient data is typically ad hoc, with the majority of the studies failing to attempt a systematic account of gradient grammaticality.

2.2.2. Competence and Performance

Generative linguistics is based on the "fundamental distinction between *competence* (the speaker-hearer's knowledge of his language) and *performance* (the actual use of the language in concrete situations)" (Chomsky 1965: 4; see also Chomsky 1995: 14–18). This definition seems to be widely shared among generative linguists and "[t]he goal of linguistic theory, under this view, is to describe the knowledge [of language], independent of (and logically prior to) any attempt to describe the role that this knowledge plays in the production, understanding, or judgment of language" (Schütze 1996: 20).

In this setting, a sentence is *grammatical* if it is generated by the grammar of the speaker i.e., it is in accordance with the linguistic knowledge that the speaker has. Grammaticality is therefore a notion that pertains to linguistic competence. Whether a sentence is *acceptable*, on the other hand, is a question about linguistic performance, it pertains to the behavior that the speaker exhibits. Linguistic judgments generated by the speaker are part of this behavior, i.e., they constitute performance data.[1]

The competence/performance dichotomy entails that acceptability judgments are not sufficient to determine the grammaticality of a sentence. The performance of a native speaker may be affected "by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attentions and interest, and errors (random or characteristic) in applying his knowledge of the language" (Chomsky 1965: 3). This means that theoretical considerations have to come into play to decide the grammaticality status of a sentence. A linguist might want to assume that certain sentences are grammatical, even though they are not accepted by native speakers (or vice versa, as discussed below). For instance if a set of examples is clearly acceptable, it can be concluded that other structurally related examples should also be generated by the grammar, even though they might not be acceptable. In such cases, the linguist's intuition about what grammars look like is more relevant than the native speaker's intuition about acceptability.

In the present thesis, we will assume that the competence/performance distinction carries over essentially unchanged to the investigation of degrees of acceptability (as opposed to the investigation of binary acceptability that characterizes mainstream linguistics). This means that we assume that some aspects of gradient data are due to factors that pertain to grammatical competence, while other aspects are due to performance effects. The decision which aspects to subsume under competence and which ones to treat as performance is ultimately a theoretical one; it cannot be settled on purely empirical grounds.

[1] It is important to note that "[i]t does not make any sense to speak of grammaticality judgments given Chomsky's definitions, because people are incapable of judging grammaticality—it is not accessible to their intuitions [...]. Linguists might construct arguments about the grammaticality of a sentence, but all that a linguistically naive subject can do is judge acceptability" (Schütze 1996: 26). The present thesis follows Chomsky's definitions and treats the terms acceptable and grammatical as distinct (contrary to the practice of many other authors, including Schütze 1996).

The assumption that the competence/performance dichotomy holds for gradient data is shared by Sternefeld (1998), who provides a detailed discussion of suboptimal (i.e., gradient) linguistic structures, from which he derives the following classification of mismatches between grammaticality and acceptability:

- **Grammaticality without Acceptability** This case arises for sentences that are grammatical, but are still rejected by native speakers on performance grounds. Well-known examples that fall in this category are garden path sentences or center embeddings, which are hard to process and therefore typically judged as unacceptable.

- **Acceptability without Grammaticality** This situation arises when a sentence is clearly ungrammatical, but is still accepted by native speakers. Examples are provided by Gibson and Thomas (1999), who show that three nested relative clause structures are just as acceptable when only two verb phrases are included instead of the grammatically required three. Gibson and Thomas (1999) provide an explanation for this effect in terms of memory limitations.

- **Overdetermined and Underdetermined Cases** This category comprises constructions where more then one (or no) grammatical rule is applicable. Examples include, for instance, subject-verb agreement in English, which seems to be subject to both conceptual number agreement and grammatical number agreement. This situation seems to manifest itself in variable judgments (Sadock 1998).

On the basis of his survey of controversial example sentences in the theoretical literature, Sternefeld (1998: 155) concludes that certain cases of gradience can be explained in terms of grammaticality/acceptability mismatches. However, the bulk of the gradient data in the literature does not seem to fall in any of the three categories. Rather, these gradient phenomena lend themselves to a *competence* description, i.e., to a description in terms of grammatical constraints. (Note also that attributing the gradience of such structures to performance factors is likely to complicate an account of the human language processor in an undesirable fashion.)

The present thesis follows Sternefeld's (1998) approach and pursues competence explanations for gradience: the judgment experiments reported in Chapter 3–5 are designed to investigate *grammatical* aspects of gradience (see Section 1.2.2). Note that this means that we are adopting an a priori position: given that no systematic performance explanation for gradience is available, we will work on the assumption that gradience is best analyzed in terms of linguistic competence.

On the other hand, there are certain extra-linguistic factors that can influence gradient linguistic judgments (measurement scale, instructions, order of presentation, field dependence, handedness, and literacy), as already mentioned in Section 1.2.2. We will assume that these influences can be factored out by applying rigorous experimental controls when gathering the

judgments. A discussion of the relevant extra-linguistic factors will be provided in the next two sections.

2.3. Factors Influencing Acceptability Judgments

"A great deal is known about the instability and unreliability of judgments" (Schütze 1996: 1), and Schütze (1996: 98–169) devotes a large part of his book to a discussion of the factors that can influence judgment behavior and engender such instability and unreliability. His conclusion is that "grammaticality judgments [...], while indispensable forms of data for linguistic theory, require new ways of being collected and used" (Schütze 1996: 1).[2] The present section considers the most relevant factors that influence acceptability judgments, with special focus on the effect of measurement scales and instructions.

2.3.1. Measurement Scales

If acceptability judgments are to be considered empirical data in the sense of experimental psychology, then the measurement scale used for judgment elicitation is of crucial importance: it determines what type of data is obtained and which mathematical (statistical) operations can be carried out on the data. Schütze (1996: 77–81) discusses the types of measurement scales that are commonly used in the experimental literature and assesses their respective usefulness for eliciting acceptability judgments. Similar overviews are provided by Bard et al. (1996) and Cowart (1997: 67–77), who make the case for the use of an interval scale for measuring acceptability (see also Lodge 1981, who argues for the use of interval scales in sociological questionnaires).

2.3.1.1. Nominal Scales

A nominal scale consists of a set of category labels representing the possible values of the property to be measured. The categories are assumed to be discrete and the only formal relation defined on the categories is equality: two stimuli can be compared as to whether or not they fall into the same category with respect to a given property. Note that no ordering relation is defined for a nominal scale, and the only mathematical operation that can be performed is counting. Hence statistical tests on nominal data have to be carried out on category frequencies.

Traditionally, linguistic examples are assigned labels like "acceptable" and "unacceptable", i.e., they are measured on a nominal scale. Such an approach assumes that acceptability is a binary notion, i.e., an individual speaker will either accept or reject an individual sentence. Under this assumption, gradience can only emerge if the judgments of a number of speakers are pooled and frequency statistics are computed.

[2] Similar issues are discussed in the literature on second language research (see Birdsong 1989; Chaudron 1983 for overviews).

2.3.1.2. Ordinal Scales

An ordinal scale has the same properties as a nominal scale, but in addition, an ordering relation is defined over the categories: stimuli can be compared in terms of their rank on the scale with respect to the measured property. However, no commitment is made as to the distance of the points on an ordinal scale, and again the only mathematical operation defined is counting, allowing frequency statistics only.

Acceptability is measured on an ordinal scale if the traditional binary categories of acceptable and unacceptable are complemented by intermediate ones. This is common practice in contemporary linguistic theory, where symbols like "?", "??", or "?*" are used in addition to the traditional "*" to record gradient acceptability judgments. This practice can be systematized by defining a consistent ordinal scale for acceptability, and much of the experimental literature on linguistic judgments has followed this practice. However, it is an open question "how many meaningful distinctions of levels of acceptability (relative or absolute) can be made" (Schütze 1996: 77). Different experimental studies have used a variety of different scales, typically consisting of three to twenty categories. An additional difficulty is that there is no agreement on the definition of the categories.

This lack of agreement is problematic, as using the right measurement scale is crucial for obtaining consistent data: "if you have too few levels, people collapse true distinctions arbitrarily, whereas if you have too many, people create spurious distinctions arbitrarily" (Schütze 1996: 78). It is conceivable that there is no ordinal scale that is optimal for all cases; the number of categories to be distinguished may vary with the linguistic phenomenon under consideration (this would explain the disagreement in the literature on which scale to use). On top of this problem, there are other difficulties with ordinal data, such as the question of how to quantify inter- and intrasubject consistency, and the fact that relative judgments can be non-transitive (Schütze 1996: 78–81).

2.3.1.3. Interval Scales

Just like an ordinal scale, an interval scale presupposes that an ordering is defined over the measured categories. In addition, a distance relation has to be defined, i.e., it has to be possible to specify the difference of any two points on the scale. Typically, an interval scale is used for properties which can be measured numerically. Mathematical operations defined on interval scales include addition and multiplication; therefore means can be calculated for the measured values and parametric statistical tests can be carried out.

Standardly, linguistic data are not measured on an ordinal scale: it is determined whether an example is more or less acceptable than another one, but not how much more or less acceptable it is. Recently, however, a number of researchers have argued that linguistic intuitions should be elicited on an interval scale using magnitude estimation, an experimental

paradigm that has been shown to yield reliable and fine-grained measurements of linguistic intuitions (Bard et al. 1996; Cowart 1997; Sorace 1992). Magnitude estimation seems particularly suitable for addressing the problems raised by the use of gradient acceptability judgments (see Section 2.2.1), and we will use magnitude estimation for all the experiments reported in Chapters 3–5. The paradigm is described in more detail in Section 2.5.

2.3.2. Instructions

The instructions used for judgment elicitation have considerable influence on the outcome of a judgment experiment. In most experiments, the speakers that function as subjects are naive, and hence likely to be unfamiliar with the linguistic concepts that they are supposed to apply in rating the stimuli. If no definitions for "grammaticality" or "acceptability" are provided, each subject will use his or her own interpretation of these concepts, and the resulting data are likely to be very noisy. Schütze (1996) observes that the majority of the studies he reviewed failed to employ adequate instructions, and hence the data they report might be confounded and have to be interpreted with caution.

In this context, Schütze (1996) refers to an experiment by Cowart that aimed to assess the role of the instructions in eliciting acceptability judgments (reported also in Cowart 1997: 55–61). This study used two types of instructions for judging the same set of sentences. The first, "intuitive", set of instructions asked subjects to base their ratings on their own reactions to a sentence, and stressed that there are no right or wrong answers. The second, "prescriptive" set of instructions evoked the scenario of an English professor marking term papers, and required subjects to judge whether a sentence would be considered right or wrong in such a context. No significant difference was found between the judgments for the two types of instructions, which leads Cowart (1997: 58) to suppose that "informants have very little ability to deliberately adjust the criteria they apply in giving judgments".

Schütze (1996) concludes that "as long as subjects are given *some* explicit set of instructions, the exact contents of those instructions might not matter a great deal, at least for some classes of sentence types" (Schütze 1996: 133). As Cowart's results show, this might be true for the instructions regrading the *criteria* subjects are supposed to apply in their judgments. However, the *rating scale* on which subjects express their judgments has been shown to be influenced by the instructions: Bard et al. (1996) found that subjects resorted to a familiar ordinal scale (a ten point scale used for marking in school), unless they were explicitly instructed not to do so. Only in this case could proper interval data be elicited. Note also that Gordon and Hendrick (1997) found that the type of instructions can have an influence on judgments of the coreference of noun phrases (see Section 5.3 for a discussion of Gordon and Hendrick's (1997) results.)

2.3.3. Subject-Related Factors

Individual differences occur in many aspects of human cognition, and have also been shown to influence acceptability judgments. A relevant individual factor is field dependence, a concept used in personality assessment. “A field dependent (FD) person fuses aspects of the world and experiences it globally, whereas a field independent (FI) person is analytical, differentiating information and experiences into components” (Schütze 1996: 177). Field dependence can be assessed using several standard tests (such as the embedded figures test), and Nagata (1989b) demonstrated that it has an influence on linguistic judgment behavior. The judgments of FI subject change with repeated exposure to the same sentence, while the judgments of FD subjects do not. A follow-up study by Cowart et al. (1998), however, failed to fully support Nagata’s (1989b) results on field dependence.

Another relevant factor is handedness, which is known to influence other aspects of linguistic behavior (e.g., sentence processing). Handedness effects in linguistic judgments are not unexpected, and indeed a study by Cowart (1989b) found effects of familial handedness on judgments of sentences with subjacency violations: right-handed speakers without left-handed relatives are more sensitive to subjacency violations (rate them as less acceptable) than right-handers that have left-handed relatives.

A contentious issue is whether linguistic training has an influence on acceptability judgments, and in particular whether linguists and non-linguists differ in their judgments. Schütze (1996: 113–122) discusses this question in some detail, and concludes that the available experimental evidence is not sufficient to establish systematic differences between the judgments of linguists and those of naive speakers. However, according to Schütze (1996), “we have enough reasons to *expect* [judgments of linguists] to be different that linguists simply ought to be excluded [as informants]” (Schütze 1996: 187).[3]

Cowart (1997: 60) concurs: “Although it might be that sustained practice can sharpen an individual’s ability to give reliable judgments, there are also reasons to suspect (as has often been suggested) that training can produce some theory-motivated bias.” Both authors conclude that only data from naive speakers should be used. Schütze (1996) deplores the fact that this suggestion is almost never followed by linguists, who “first consult their own intuitions (one cannot find a more biased subject than the investigator), then their colleagues in the next office (almost as biased), and if they are really ambitious, perhaps a couple of their students (not exactly objective either, since students are likely to know which results their professors are hoping for and would like to gain their favor)” (Schütze 1996: 187).

[3] Cowart (1997: 60) goes a step further and argues that, while it is possible in principle to experimentally establish the influence of linguistic training on judgments, relevant experimental studies are unlikely to become available, as they are very difficult to carry out for practical reasons. (Such experiments would involve a standardized linguistic training program to be administered to a group of naive subjects, while monitoring the effects on their judgment behavior.) The lack of relevant studies “makes data obtained from expert informants particularly difficult to interpret”, hence Cowart gives preference to “evidence that does not rely on expert skills of unknown reliability” (Cowart 1997: 60).

2.3.4. Task-Related Factors

Measurement scale and instructions are important task-related factors in acceptability judgments, as argued in sections 2.3.1 and 2.3.2. Another task-related factor discussed by Schütze (1996) is order of presentation. In experiments on the acceptability of participle adjuncts, Greenbaum (1973, 1976) found that a sentence is judged less acceptable if it is presented at the first position of a list of sentences. Order effects are also reported by Greenbaum and Quirk (1970), and Schütze (1996: 134) concludes: "[c]learly, then, sentence order should be controlled for, either by randomization or counterbalancing". Cowart (1997: 94) agrees and points out that "the informant's state of mind may well change in relevant ways as she proceeds through the [acceptability judgment] questionnaire. Fatigue, boredom, and response strategies the informant may develop over the course of the experiment can have differing effects on sentences judged at various points in the entire procedure".

A well-established factor in judgment behavior is repetition. The repetition effect and its interaction with other factors (such as field dependence) has been examined extensively by Nagata (1987, 1988, 1989a,b,c). These results show that repetition within a short interval leads to lower acceptability ratings, while repetition after a long interval (four months) has no significant influence on judgments. Schütze (1996) notes that the repetition effect also manifests itself in what is known as "linguists' disease", i.e., the phenomenon that one's acceptability judgments become increasingly blurred and uncertain when one ponders long enough over many examples of the same type. (This is another argument against relying on judgments provided by linguists.)

Another potential influence is mode of presentation: several studies have looked at the differences induced by the visual or auditory presentation of sentences. It has been suggested that the more formal mode of written presentation should lead to more stringent judgments, but Schütze (1996: 147–149) concludes that the literature provides no firm evidence for this. Finally, a number of studies have investigated the so-called anchoring effect: if a sentence is judged as part of a set of severely unacceptable sentences it will receive a higher rating than if it is part of a set of acceptable (or mildly unacceptable) stimuli. However, Cowart (1994) demonstrated that, while the anchoring of experimental stimuli influences the absolute ratings, it does not seem to affect relative judgment patterns.

2.4. Eliciting Reliable Acceptability Judgments

As we saw in Section 2.3, acceptability judgment behavior is influenced by a diverse number of factors, both task-related and subject-related. Unless they are properly controlled for, these factors can introduce a considerable amount of variance into the data, which leads Schütze (1996) to urge the use of experimental methods to obtain reliable judgments: "considerable care and effort must be put into the elicitation of grammaticality judgments if we are to stand a chance

of getting consistent, meaningful and accurate results" (Schütze 1996: 171). This is particularly true for gradient judgments, as argued in Section 2.2.1. In what follows, we discuss Schütze's (1996) and Cowart's (1997) recommendations and describe how they are implemented in the experimental methodology applied in the present thesis.

2.4.1. Materials

To minimize potential biases, Schütze (1996) suggests a number of basic controls that should be applied when designing the sentence materials for an acceptability judgment experiment. Detailed recommendations on how to construct sentence materials are also provided by Cowart (1997), as part of a comprehensive introduction to the design of acceptability judgment experiments.

Presentation order as a potential confounding factor should be avoided by counterbalancing the order of the stimulus sentences across different subjects. This can also be achieved by randomizing the order of the materials for each subject, which is the option used in the experiments reported in this thesis.

Also, the stimulus set should not contain substantially more acceptable than unacceptable sentences (or vice versa), as otherwise subjects might fall into a yea-saying or nay-saying mode, or develop expectations about the stimuli that might bias their responses. We adhere to this criterion by selecting the stimulus set and the fillers such that the number of acceptable items roughly matches the number of unacceptable items.

Another potential confounding factor is the lexicalization of the stimulus sentences. Instead of testing individual sentences, an experiment should investigate sentences types, where each sentence type is represented by several lexicalizations. In choosing the lexicalizations, we have to take frequency into account, as the frequency of a lexical item can influence judgment behavior. In the present thesis, this problem is addressed by balancing the experimental materials for frequency, based on corpus counts for the relevant lexical items.

Furthermore, Schütze (1996) recommends the use of contextualized experimental sentences, as "there are numerous ways that context can influence grammaticality, from bringing out rare word meanings to priming certain parsing procedures" (Schütze 1996: 185). Such effects should be controlled for, i.e., "a supporting pragmatically related context should always be provided" (Schütze 1996: 185). Otherwise subjects will make up their own contexts, thus potentially increasing inter-subject variance in the ratings. All the experiments in Chapter 4 use contextualized stimuli, but also include a null context condition as a control. The results indicate that for isolated sentences, subject seem to make minimal contextual assumptions (the judgment patterns in a null context match the ones in an information structurally neutral context, i.e., an all focus context). This justifies the use of isolated stimuli in the experiments in Chapter 3.

Sentences that might trigger processing problems should be excluded from the test

materials, as they are likely to confound the acceptability ratings (e.g., garden path sentences and center embeddings are rated unacceptable, as demonstrated by Marks (1968) and Warner and Glass (1987)). To our knowledge no such materials are contained in the data sets used for the experiments in Chapters 3–5.

To obtain maximally fine-grained results, the stimulus set should consist of minimal pairs, i.e., the sentences should "be matched as closely as possible on as many features as possible, including semantic plausibility" (Schütze 1996: 186). This suggestion is adhered to in all our experimental designs.

2.4.2. Procedure

Once necessary steps have been taken to reduce confounds in the sentence materials, the next aim should be to minimize potential biases in the procedure of gathering judgments. Again Schütze (1996) and Cowart (1997) provide a set of very useful recommendations, which will be summarized in the following.

Schütze (1996) considers the selection of subjects the most important procedural issue. "If it is the competence of normal native speakers that we claim to be investigating, we need to study random samples of normal native speakers" (Schütze 1996: 186f). In particular, linguists should be excluded as informants, as their judgments are likely be confounded by theoretical bias (see Sections 2.3.3 and 2.3.4). "If linguists wish to live up to scientific standards of data validity, it is time for them to abandon the convenient fiction that data is never further away than their own minds" (Schütze 1996: 187). We follow this recommendation: all the experiments in Chapters 3–5 use naive native speakers as subjects, i.e., speakers that have had no prior linguistic training.

Furthermore, the number of subjects used has to be large enough so that statistical test can be carried out on the data. (This is another argument against the use of linguists as informants, as normally only small numbers of linguists are available.) This recommendation is adhered in all experiments reported in the present thesis.

Potentially relevant individual differences (see Section 2.3.3) should be recorded on a questionnaire that accompanies the acceptability judgment experiments. This allows us to test for an influence of these factors on the judgment data. Cowart (1997: 168f) gives an example questionnaire on individual differences that includes sex, education, age, handedness, language variety, and linguistic training. All our experiments include a pre-test questionnaire that is based on Cowart's (1997) recommendations. In some cases the results from this questionnaire are relevant for the evaluation of the experimental data, e.g., in Experiments 1–3, where we make use of data on the dialect region that subjects belong to.

As discussed in Section 2.3.2 the instructions given to subjects are likely have an important influence on the reliability of the judgment results. In particular, as Schütze (1996) points out, "one cannot hope for the terms *grammatical* or *acceptable* to have their intended

meanings for naive subjects" (Schütze 1996: 188). He argues that the instructions should be as specific as possible in defining these terms, preferably making reference to relevant examples. This suggestion is implemented in the experiments reported in the present thesis: we use a set of instructions that is based on recommendations in the magnitude estimation literature (see Lodge 1981). In these instructions, the concept of "acceptability" is defined by example.

Schütze (1996) gives no clear recommendation as to the rating scale that should be used. He holds that both relative and absolute ratings can be appropriate, depending on the issue under investigation. Recent studies, however, favor the use of an interval scale based on the magnitude estimation paradigm. Magnitude estimation has been shown to yield reliable and maximally fine-grained judgment data (Bard et al. 1996; Cowart 1997), while avoiding the problems with conventional ordinal scales (see Sections 2.3.1). In particular, Sorace (1992) demonstrated that magnitude estimation can detect acceptability differences that go unnoticed if an ordinal scale is used. The present thesis uses the magnitude estimation paradigm for all experiments.

To ensure that subjects apply instructions and rating scale as intended, the judgment experiment should be preceded by a series of warm-up trials, preferably involving sentences similar to the ones used as experimental materials. All our experiments include two warm-up phases: the first one is designed to familiarize the subjects with the concept of magnitude estimation, the second one allows them to practice magnitude estimation on linguistic stimuli similar to the ones used in the actual experiment..

In designing the materials for a judgment experiment, it is important to use a sufficient number of filler sentences, i.e., to present the experimental items interspersed in a list of sentences that are unrelated to the constructions under investigation. The fillers serve to prevent subjects from becoming aware of the purpose of the experiment (as this might bias their judgments). Also, the fillers allow the experimental items to be anchored, thus making sure that subjects make proper use of the rating scale (fillers should cover the whole acceptability range, see Cowart 1994). In all experiments reported in the present thesis, about half of the sentences in each stimulus set are fillers.

2.4.3. Evaluation

A certain amount of variance will remain in the experimental data, even if all necessary controls are applied. This variance could either be due to chance or could result from the experimental manipulation, i.e., from a factor that the experiment is meant to investigate (e.g., the violation of a certain grammatical constraint). In the latter case, the effect (e.g., a difference in acceptability) is significant, in the former case non-significant. The only way of determining the significance of an effect is by performing statistical tests on the data, and so Schütze's (1996) most important recommendation the use of statistical methods, a suggestion that "linguists consistently ignore" (Schütze 1996: 195). This point is particularly important if degrees of acceptability are

investigated: pure intuition is not sufficient for determining whether small differences in acceptability are reliable or not. (Cowart (1989a, 1997) demonstrates this point with respect to extraction for picture NPs.) We adhere to this recommendation regarding evaluation in all experiments reported in Chapters 3–4. Standard experimental statistics (analysis of variance and associated post-hoc tests) are used to determine significant differences in acceptability.

Schütze (1996: 186–201) also considers the problem of inconsistencies in judgments, i.e., how to interpret disagreements between speakers or changes over time in the ratings of a single speaker. Experimental evidence presented by Cowart (1997) shows that the overall judgment pattern for a given structure can be highly stable within a group of speakers, while at the same time, the judgments of individual speakers show considerable variance. Cowart concludes that, similar to other types of behavioral data, linguistic judgments seem to exhibit a certain amount of random variance around a stable mean, which he takes as a strong argument for collecting judgment data experimentally.

2.5. Magnitude Estimation

The present thesis relies crucially on subtle linguistic intuitions, viz., on judgments of the relative acceptability of competing linguistic structures. Such relative acceptability judgments should be measured experimentally, since the informal elicitation technique traditionally used in linguistics is unlikely to be reliable for such data, as argued extensively in Sections 2.2–2.4. A suitable experimental paradigm is magnitude estimation (ME), a technique standardly applied in psychophysics to measure judgments of sensory stimuli (Stevens 1975). The magnitude estimation procedure requires subjects to estimate the perceived magnitude of physical stimuli by assigning values on an interval scale (e.g., numbers or line lengths) proportional to stimulus magnitude. Highly reliable judgments can be achieved in this way for a whole range of sensory modalities, such as brightness, loudness, or tactile stimulation (for an overview, see Stevens 1975).

The ME paradigm has been extended successfully to the psychosocial domain (see Lodge 1981 for a survey) and recently Bard et al. (1996), Cowart (1997), and Sorace (1992) showed that linguistic judgments can be elicited in the same way as judgments for sensory or social stimuli. Unlike the five- or seven-point scale conventionally employed in the study of psychological intuition, ME allows us to treat linguistic acceptability as a continuum and directly measures acceptability differences between stimuli. ME's use of an interval scale means that parametric statistical tests can be applied for data analysis.

ME has been shown to provide fine-grained measurements of linguistic acceptability, which are robust enough to yield statistically significant results, while being highly replicable both within and across speakers. ME has been applied successfully to phenomena such as auxiliary selection (Bard et al. 1996; Sorace 1992, 1993a,b; Sorace and Cennamo 2000; Sorace

and Vonk 1998), coordination and binding (Cowart 1997), resumptive pronouns (McDaniel and Cowart 1999), *that*-trace effects (Cowart 1997), compounding (McDonald 1995), extraction (Cowart 1997; Keller 1996a,b), and selectional restrictions (Lapata, McDonald, and Keller 1999).

The ME procedure for linguistic acceptability is analogous to the standard procedure used to elicit judgments for physical stimuli. Subjects are presented with a series of linguistic stimuli, and have to respond by assigning a value to each stimulus proportional to the acceptability they perceive. Several different modalities can be used for expressing the response values,[4] but previous studies tended to use either numeric values (e.g., Sorace 1992, 1993a,b) or line lengths (e.g., McDonald 1995). Both modalities suffer from specific drawbacks. Numeric judgments tend to exhibit an integer bias, as subjects prefer to use integers instead of making estimates in the range of decimal numbers. Line drawing, on the other hand, has the problem of physically restricting the range of subjects' responses (as the space provided on a screen or on a piece of paper is limited). In many cases, a regression bias is found for line drawing, i.e., subjects commonly draw unproportionally short lines for items at the upper end of stimulus range.

Bard et al. (1996) used a cross-modal matching paradigm to show that ME data are consistent when elicited cross-modally, i.e., using both numeric values and line lengths as response modalities. Similar results are reported by Cowart (1997). We conclude that the choice of response modality is essentially arbitrary, and decided to use the numeric modality for the experiments in this thesis, as this facilitates data collection and evaluation.

2.6. An Introduction to Optimality Theory

Standard Optimality Theory deviates from more traditional linguistic frameworks in that it assumes grammatical constraints to be (a) universal, (b) violable, and (c) ranked. Assumption (a) means that constraints are maximally general, i.e., they contain no exceptions or disjunctions, and there is no parameterization across languages. Highly general constraints will inevitably conflict; therefore assumption (b) allows constraints to be violated, even in a grammatical structure, while assumption (c) states that some constraint violations are more serious than others. While, according to (a), the formulation of constraints remains constant across languages, the ranking of the constraints can differ between languages, thus allowing us to account for crosslinguistic variation.

In an OT setting, a structure is grammatical if it is the *optimal* structure in a set of candidate structures. Optimality is defined via constraint ranking: the optimal structure violates the least highly ranked constraints compared to its competitors. The number of violations plays

[4] An overview of response modalities is given by Lodge (1981: 24ff), who also discusses the validation of ME results via cross-modal matching.

Table 2.1: Constraint profile for direct object extraction (simplified from Legendre et al. 1995: (22a))

	$[Q_j$ [think$_{CP}$ $[x_j]]]$	SUBCAT	BAR4	BAR3	BAR2	*t
a.	what$_j$ do [you [think [he [said t_j]]]]	*		*		*
b.	what$_j$ do [you [think [t_j that [he [said t_j]]]]]				**	**
c.	what$_j$ do [you [think [that [he [said t_j]]]]]		*			*

a secondary role; if two structures violate a constraint with the same rank, then the number of violations incurred decides the competition. OT therefore deviates from traditional grammatical frameworks in that the grammaticality of a sentence is not determined in isolation, but in comparison with other possible structures. Note that there is no inherent restriction on the number of optimal candidates for a given candidate set; more than one candidate may be optimal if several candidates share the same constraint profile, i.e., if they incur the same constraint violations.

We will illustrate how OT works with a simple example taken from an account of *wh*-extraction by Legendre, Wilson, Smolensky, Homer, and Raymond (1995). Our example deals with extraction from direct objects in English. Legendre et al. (1995) assume that the following constraints govern extraction: SUBCAT, which states that the subcategorization requirements of the verb have to be met; *t, which disallows traces (i.e., movement); and BAR*n*, which rules out movement that crosses more than *n* barriers (for a definition of barrier, see Legendre et al. 1995). For English, the assumption is that these constraints are ranked as follows:

(2.1) SUBCAT $\gg$ BAR4 $\gg$ BAR3 $\gg$ BAR2 $\gg$ *t

This means that a violation of SUBCAT is more serious than a violation of BAR4, which in turn is more serious than a violation of BAR3, etc.

A crucial assumption in OT is that all candidate structures (syntactic representations) that take part in a grammatical competition are generated from a common input, assumed to be a predicate argument structure by Legendre et al. (1995). The input structure specifies the verb and the arguments of the verbs, plus operators and scope relations that might be present. (Section 7.1.1 sets out the assumptions that the present thesis makes about the input in more detail.) As an example, consider the first line of Table 2.1: This input contains the verb *think* (subcategorizing for a CP complement) and specifies that its argument has to contain a syntactic variable x_j which is in the scope of a question operator Q_j. Such an input has to be realized by a *wh*-question. (For a discussion of the problem of input representations, see Section 7.1.1.)

Possible realizations of this input are the candidates (a)–(c) in Table 2.1. These candidates violate different constraints, as indicated by the asterisks in Table 2.1. For example,

candidate (a) violates SUBCAT (as the verb takes an IP complement, instead of a CP complement), *t (due to the moved *wh*-element it contains), and BAR3 (because the movement crosses three barriers).

The *optimal* structure in a candidate set is computed as the structure that violates the least highly ranked constraints. As an example, consider the competition between candidates (a) and (c): (a) violates SUBCAT, while (c) violates BAR4. According to the constraint hierarchy in (2.1), SUBCAT is ranked higher than BAR4, which means that candidate (c) wins the competition. Note that all the other constraints that are violated by either of the candidates are not taken into account in determining the winner. Only the most highly ranked constraint on which the two candidates differ matters for the constraint competition (*strict domination* of constraints). Two candidates differ on a constraint if one candidate violates that constraint more often than the other one (e.g., (a) violates SUBCAT once, while (b) violates it zero times).

In Table 2.1 the optimal candidate is (b): It wins against (c), as it violates BAR2 instead of BAR4. The additional trace that (b) contains allows it to avoid crossing four barriers at once. This means that (b) incurs two violations of *t (instead of just one). However, this is not relevant to the competition with (c), due to strict domination. (Note that (a) would win if the input contained *think* subcategorizing for an IP.)

Another important aspect of OT can also be illustrated using the extraction example: In OT, crosslinguistic variation can be accounted for by *constraint re-ranking*. Assume that there is an additional constraint *Q, which disallows empty question operators. For English, the ranking *Q $\gg$ *t holds. This means that questions are formed by movement of *wh*-elements, while in-situ *wh*-elements, which have to be bound by the Q operator, are ungrammatical. Chinese, on the other hand, exhibits the opposite ranking *t $\gg$ *Q, i.e., the use of an empty question operator is preferred to the use of a trace. This explains why in Chinese, *wh*-elements remain in situ in direct object extractions, where the *wh*-element is bound by the Q operator. English, on the other hand, requires *wh*-movement in such configurations, as illustrated by the example in Table 2.1.

2.7. Conclusions

This chapter presented the background for both the experimental (Chapters 3–5) and the theoretical part (Chapters 6 and 7) of this thesis. The main goal was to provide an overview of the methodological issues related to linguistic judgments. We discussed the practice of using acceptability judgments in linguistics and pointed out potential problems with this practice in general, and with its application to gradient data in particular. We also dealt with the competence/performance dichotomy that underlies linguistic theory and argued that it applies to gradient judgments essentially unchanged.

Following Schütze (1996), we reviewed the literature on acceptability judgments and

concluded that there is a strong case for the use of experimental methods for eliciting judgment data. A multitude of factors that potentially affect judgment behavior has been identified, and the conventional intuitive approach to judgment collection is clearly inadequate to control for these factors. Experimental procedures have to be applied to minimize potential biases in judgment elicitation, and experimental statistics has to be employed to establish the significance of observed differences in acceptability.

Based on this premise, we reviewed the recommendations by Schütze (1996) and Cowart (1997) on how to collect reliable judgment data, and discussed the implementation of these recommendations in the experiments reported in this thesis. We also gave an overview of the magnitude estimation paradigm that is used throughout the thesis.

Finally, we presented an introduction to Optimality Theory, the competition-based grammatical framework that guides both the experimental and the theoretical investigations in the remainder of this thesis.

Chapter 3

Gradient Grammaticality out of Context

This chapter presents a series of experiments that establish a number of general properties of gradient linguistic judgments. The experiments deal with unaccusativity, extraction, binding, and word order, and the aim is to investigate how constraint ranking, constraint type, and constraint interaction determine the degree of acceptability of a given linguistic structure.

The experimental findings indicate that two fundamental properties of linguistic constraints are responsible for gradience in grammar. Firstly, constraints are ranked, in the sense that some constraint violations lead to a greater degree of unacceptability than others. Secondly, constraint violations are cumulative, i.e., the degree of unacceptability of a structure increases with the number of constraints it violates.

The results reported in this chapter also indicate that two constraints types can be distinguished experimentally: soft constraints lead to mild unacceptability when violated, while hard constraint violations trigger serious unacceptability. Crosslinguistic studies lead to the hypothesis that only soft constraints are subject to crosslinguistic variation, while hard constraints are immune to crosslinguistic effects.

This hypothesis, as well as the interaction of soft and hard constraints with context, will be subject to further experimental study in Chapter 4.

3.1. Introduction

The aim of the present thesis is to provide an experimentally motivated model of degrees of grammaticality. In Section 1.2.2 we distinguished between linguistic and extra-linguistic factors that influence gradient judgments, and in Sections 2.3 and 2.4 we argued that extra-linguistic influence can be factored out with by applying rigorous experimental controls in the collection of gradient judgments.

Linguistic factors can be further subdivided into competence and performance factors, and in Section 2.2.2 we argued that this competence/performance distinction carries over essentially unchanged to the study of degrees of grammaticality (as opposed to the study of binary grammaticality). The present thesis will pursue competence explanations of gradience, i.e., in the absence of systematic performance explanations, we will assume that gradience pertains to the linguistic knowledge of the speaker, traditionally considered the domain of linguistic theory.

The experiments presented in Chapters 3 and 4 are designed to address a set of fundamental questions regarding competence aspects of gradience in grammar. In the following section, we provide a brief outline of these questions.

3.1.1. Constraints

As the basis for investigating questions regarding gradient linguistic structures, we have to establish a set of linguistic constraints that allow us to formulate these questions.

We use the term constraint in a fairly theory-neutral sense, referring to a principle or rule of grammar that can be either satisfied or violated in a given linguistic structure. While we draw some of our constraints from the existing theoretical literature, we generally adopt a notion of constraint that is essentially descriptive. By this we mean that whether a constraint is violated or not can be read off the surface string of a given sentence (e.g., the subject precedes the object), as opposed to being the consequence of underlying theoretical constructs (e.g., the subject has moved to specifier position).

In this sense, our use of the term linguistic constraint diverges from its use in current linguistic frameworks such as Optimality Theory (OT; Prince and Smolensky 1993, 1997; see Section 2.6 for an overview), which rely on a theory-driven notion of constraint. We opt for descriptive constraints as these allow us to formulate our results in a manner that is largely theory-neutral. This is desirable as we are mainly interested in questions pertaining to the behavior of constraints (constraint type, constraint ranking, constraint interaction, detailed below), rather than in the constraints proper.

Nevertheless, the experimental results presented in Chapters 3 and 4 make a contribution to linguistic theory. Each experiment investigates the influence of a set of constraints on the acceptability of a certain linguistic structure. By demonstrating such an influence (or its absence), our experimental data contribute to settling data disputes in the theoretical linguistic literature. The underlying assumption is that such data disputes are the results of the informal data collection techniques employed in theoretical linguistics, which are not well-suited to investigate the behavior of gradient linguistic data (as argued in Sections 1.2 and 2.4).

3.1.2. Constraint Ranking

The first question to be addressed in the present chapter concerns constraint ranking. Our aim is to provide evidence for the fact that linguistic constraints are ranked, i.e., differ in their relative importance.

To investigate this question experimentally, we employ an operational definition of constraint ranking based on the relative unacceptability caused by a constraint violation; the higher the degree of unacceptability caused by the constraint violation, the more important the constraint. In other words, a constraint C_1 is ranked higher than a constraint C_2 if a violation of C_1 leads to a higher degree of unacceptability than a violation of C_2.

This definition of constraint ranking differs from the one standardly employed in Optimality Theory. In OT, constraint ranking is merely a tool for formalizing constraint competition; no direct correspondence between constraint ranks and degrees of acceptability is assumed. Constraint rankings are used to determine the optimal candidate in a set of candidate structures. This optimal candidate is predicted to be grammatical; no predictions are made about suboptimal constraints and their degree of ungrammaticality.

We will make crucial use of evidence on constraint ranking in developing a model of gradience in Chapter 6. The task of this model will be to predict the degree of grammaticality of a given structure from the ranks of the constraints the structure violates.

3.1.3. Constraint Types

The second question we address in this chapter deals with constraint types. The aim is to determine if gradience affects all linguistic constraints in the same way, or if it is restricted to certain constraints, while other constraints trigger binary acceptability judgments.

This leads to the more general question whether gradient data can provide criteria for a classification of constraints into constraint types. One criterion for such a classification is whether a given constraint triggers gradience or not. Other criteria include the interaction of gradience with other linguistically important factors, such as crosslinguistic variation or context. The present chapter will deal with crosslinguistic variation, while Chapter 4 will investigate the interaction of gradient grammaticality and context.

A classification of constraints into types is important for the design of a model of gradience (the subject of Chapter 6). Such a model should either incorporate the classification as one of its fundamental assumptions, or it should make it possible to derive the classification from more fundamental properties of the model. In particular, the model should predict how the type of a constraint affects its behavior with respect to, for instance, crosslinguistic variation and context effects.

3.1.4. Constraint Interaction

The third question we address concerns the interaction of constraints. By constraint interaction we mean the behavior of structures that incur multiple constraint violation.

Again, we rely on an operational definition: the interaction of two constraints can be determined by investigating the degree of unacceptability of a structure that violates both constraints, and comparing it to the degrees of unacceptability of structures that violate only one of the two constraints. An accurate picture of constraint interaction can be built up by investigating structures that violate constraints of different ranks and types.

Experimental data on constraint interaction make it possible to distinguish between competing accounts of constraint interaction. Relevant theoretical proposals include OT's principle of strict domination, which states that the highest ranking constraint on which two structures conflict is crucial for deciding which of the structures is optimal (see Section 2.6). Strict domination entails that the violation of a constraint *C* cannot be compensated by any number of violations of constraints that are lower ranked than *C*. This means that there is no ganging up of multiple lower ranked constraints against a higher ranked constraint.

Other forms of constraint interaction are conceivable. A simple alternative approach would be the summation of constraint violations: here, the degree of ungrammaticality of a structure is computed from the sum of the individual constraint violations is incurs. Based on our operational definitions of constraint ranking and constraint interaction, we can compare the prediction of proposals such as strict domination or the summation of violations.

Experimental results on constraint interaction are crucial for developing the model of gradience in grammar proposed in Chapter 6, as well as for evaluating competing models of gradience that were proposed in the literature, such as the markedness model (Müller 1999) and the Gradual Learning Algorithm (Boersma 1998; Boersma and Hayes 2001; Hayes 2000).

3.1.5. Coverage

The present thesis attempts to make maximally general claims about gradient structures with respect to constraint ranking, constraint type, and constraint interaction. The experimental studies are designed to cover all major grammar modules as standardly assumed in a syntactic framework such as Government and Binding Theory (GB; Chomsky 1981, 1986), viz., Theta Theory, Movement Theory, X-bar Theory, and Case Theory (as classified in Haegeman 1994). Furthermore, we draw on experimental data from a variety of languages (English, German, and Greek). One phenomenon (word order) serves as a case study; this phenomenon will be investigated in considerable detail, and in more than one language. The data on word order allows us to formulate crosslinguistic claims, and will be utilized for an extensive test of our model of gradience (see Chapter 7).

Table 3.1 gives an overview of the syntactic phenomena investigated in this thesis, and

Table 3.1: Syntactic modules covered by the experimental data

module	phenomenon	exp.	language	factors
Theta Theory	unaccusativity, unergativity	1, 2, 3	German	verb class, animacy, telicity
Movement Theory	extraction	4, 9	English	verb class, referentiality, definiteness, inversion, resumptive pronouns, agreement
Binding Theory	exempt anaphors	5	English	verb class, referentiality, definiteness, intervening binder
X-bar Theory	gapping	7, 8	English	verb frame, remnant type, subject-predicate interpretation, simplex sentence
Case Theory	word order	6, 10, 11, 12	Greek, German	case marking, pronominalization, verb position, clitic doubling, accent placement

also lists the relevant experimental factors and the grammar modules and languages covered.

3.1.6. Acceptability Marks

The use of acceptability marks like "?" and "*" is problematic for gradient data, as discussed in Section 2.3.1.2. For expository purposes, however, we will supply acceptability marks for the example sentences cited in this thesis. While this goes against the general approach of the thesis (viz., relying on experimental data instead of on intuitive acceptability ratings), it was felt that omitting acceptability marks would make the argumentation hard to follow.

Therefore, the following convention will be adopted throughout the present chapter and Chapter 4. We will use "*" to mark a sentence that incurs at least one violation of a hard constraint, while "?" will be used to indicate at least one violation of a soft constraint. Sentences will be without acceptability mark if they do not incur violations, or if their acceptability status is unclear (and has to be settled experimentally). The meaning of the hard/soft distinction will be come clear in the course of the present chapter. Examples from the literature are reported with their original acceptability marks; the meaning of these marks might not correspond to the conventions adopted in this thesis.

3.2. Experiment 1: Effect of Verb Class on Unaccusativity and Unergativity

We start our investigation of gradience in grammar with an experiment on constraint types. The phenomenon under investigation is unaccusativity/unergativity, as manifested in auxiliary selection and impersonal passive formation in German. It has been proposed that unac-

cusative/unergative verbs can be classified into two types, core and peripheral verbs, based on their crosslinguistic behavior. The present experiment will provide support for this classification, combining evidence from gradient judgments with evidence from crosslinguistic variation. (The experiment deals with dialect variation, which we consider an instance of crosslinguistic variation.)

We will argue that two types of constraints, soft and hard constraints, underlie the distinction between core and peripheral verbs. This classification will form the basis for Experiments 4–6 in the remainder of this chapter, where the investigation of constraint types is extended to additional linguistic phenomena. The soft/hard distinction also underpins Experiments 7–12 in Chapter 4, where the investigation of context effects will provide additional support for the hard/soft dichotomy.

3.2.1. Background

Central to Sorace's (2000) account of unaccusativity and unergativity in western European languages is a classification of intransitive verbs into a set of semantic classes (which will be discussed in more detail in Section 3.2.1.1). These verb classes are organized in a hierarchy as follows:

(3.1) **Auxiliary Selection Hierarchy**

change of location	selects BE (least variation)
change of state	
continuation of state	
existence of state	
uncontrolled process	
controlled process (motional)	
controlled process (non-motional)	selects HAVE (least variation)

The Auxiliary Selection Hierarchy forms the basis for the distinction between *core* and *peripheral* verbs. Core unaccusative verbs reside at the top of the hierarchy, and select the equivalent of BE as their auxiliary. Core unergative verbs are located at the bottom of the hierarchy, and select the equivalent of HAVE as their auxiliary. As we move towards the center of the Auxiliary Selection Hierarchy, the verb classes become more and more peripheral. Peripheral verbs are subject to crosslinguistic differences and exhibit gradient auxiliary selection preferences.

Sorace (2000) also observes that the auxiliary selection behavior of peripheral verbs is influenced by non-syntactic factors such as animacy and telicity. Animacy effects can be tested by comparing the auxiliary preference for a given verb with animate and inanimate subjects. Telicity effects emerge if the auxiliary preference of a verb can be modified by adding a telic or atelic adverbial.

Sorace's (2000) classification of unergative and unaccusative verbs is based on judgment experiments for Italian (Bard et al. 1996; Sorace 1992, 1993a), French (Sorace 1993b), and Dutch (Sorace and Vonk 1998). Dialect variation has been investigated by Sorace and Cennamo (2000), who deal with auxiliary selection in Paduan. Other relevant experimental work includes the acquisition study by van Hout, Randall, and Weissenborn (1993) and Bard, Frenck-Mestre, Kelly, Killborn, and Sorace's (1999) comparison of auxiliary selection judgments with real-time measurements such as eye tracking data. To our knowledge, there are no previous experimental studies of auxiliary selection and impersonal passive formation in German.

3.2.1.1. Verb Classes

Change of Location Verbs denoting a change of location have a strong telic component and are classified as core unaccusatives by Sorace (2000). This classification seems to be crosslinguistically valid, and also extends to German, where verbs in this class select the auxiliary *sein* "be". Class members include the verbs *kommen* "come", *flüchten* "flee", *abreisen* "depart", and *entkommen* "escape", as illustrated in (3.2).[1]

(3.2) Der Gefangene ist/*hat schnell entkommen.
the prisoner is/has quickly escaped
"The prisoner escaped quickly."

The auxiliary selection behavior of change of location verbs is stable even if they are detelicized, as in (3.3); the alternative auxiliary *haben* "have" is seriously unacceptable. This confirms the status of these verbs as core unaccusatives.

(3.3) Es sind/*haben stundenlang Gefangene entkommen.
EXPL are/have for hours prisoners escaped
"Prisoner escaped for hours."

Change of State Verbs in this class denote a change of state other than a change of location. Change of state verbs can be telic, such as *versterben* "die" or *verschwinden* "disappear", or they can denote a gradual change of state, e.g., *wachsen* "grow" or *steigen* "increase". Both types select *sein* in German:

(3.4) a. Das Kind ist/*hat schnell gewachsen.
The girl is/has quickly grown
"The girl grew quickly."

b. Der Großvater ist/*hat unerwartet verstorben.
the grandfather is/has unexpectedly died
"The grandfather died unexpectedly."

Note that change of state verbs are not sensitive to detelicization:

[1] On the use of acceptability marks in this chapter, see Section 3.1.6.

(3.5) Die Temperatur ist/*hat drei Stunden lang gestiegen, dann ist/*hat sie wieder gefallen.
the temperature is/has three hours long risen then is/has it again fallen
"The temperature fell for three hours, then it rose again."

Similar construction in Dutch allow both auxiliaries (van Hout et al. 1993).

Continuation of State The verbs in this class are stative; the denote the continuation of a pre-existing state. Examples include *überleben* "survive", *dauern* "last", *verweilen* "stay", and *verharren* "persist". These verbs are unergative in German; they prefer the auxiliary *haben* "have":

(3.6) Der Wanderer ?ist/hat kurz verweilt.
the hiker is/has briefly stayed
"The hiker stayed briefly."

Verbs of this class do not seem to be sensitive to detelicization:

(3.7) a. Der Wanderer ?ist/hat auf dem Rastplatz verweilt.
the hiker is/has at the resting place stayed
"The hiker stayed at the resting place."

b. Der Wanderer ?ist/hat eine lange Zeit verweilt.
the hiker is/has a long time stayed
"The hiker stayed a long time."

Note, however, that the alternative auxiliary *sein* is not completely unacceptable with verbs of this class (see (3.6) and (3.7)), which points to the fact that the class member as peripheral unergatives. This is also evidenced by the observation that some verbs of continuation of state prefer *sein*:

(3.8) Der Wanderer ist/*hat kurz geblieben.
the hiker is/has briefly stayed
"The hiker stayed briefly."

This includes *bleiben* "remain" and its derivatives (*zurückbleiben* "stay behind", *dableiben* "stay put", etc.). Sorace (2000) points out that "remain" type verbs also show exceptional auxiliary selection behavior in French and Dutch.

Existence of State Verbs in this class denote the existence of a state. This can either be a concrete physical state in verbs like *existieren* "exist", *bestehen* "be the case", or *sein* "be", or a psychological state like in verbs such as *scheinen* "seem", *gefallen* "please", or *ausreichen* "suffice". The first category also includes verbs denoting the maintenance of a position like *sitzen* "sit", *stehen* "stand", *hocken* "squat", or *knieen* "kneel". These positional verbs exhibit gradience, i.e., they allow both auxiliaries to a certain extent:[2]

[2]In English, maintenance of position verbs such as *sit* or *kneel* also have an assume position reading (Levin and Rappaport Hovav 1995). This reading is not available for the corresponding verbs in German, which

(3.10) Die Betende ?ist/hat würdevoll gekniet.
The praying person is/has dignified kneeled
"The praying person kneeled with dignity."

The gradient auxiliary selection behavior seems to be preserved even if the verb is detelicized:

(3.11) a. Die Betende ?ist/hat auf dem Beichtstuhl gekniet.
The praying person is/has on the confessional kneeled
"The praying person kneeled on the confessional."
b. Die Betende ?ist/hat stundenlang gekniet.
The praying person is/has for hours kneeled
"The praying person kneeled for hours."

Psychological state verbs like *scheinen* "seem" or *reichen* "suffice" fail to show gradience and select *haben*. Other existence of state verbs such as *existieren* "exist" also select *haben*. An exception is *sein* "be", which also denotes existence, but selects *sein* as its auxiliary.

The heterogeneous auxiliary selection pattern in this class confirms that existence of state verbs are peripheral in the Auxiliary Selection Hierarchy (see (3.1)). Another point in case is the fact that positional verbs exhibit gradient auxiliary selection behavior, i.e., they allow both auxiliaries, at least to a certain extent. Note that dialect variation has been observed for the existence of state class; this will discussed in Section 3.2.1.3.

Uncontrolled Process The verbs in this class share the property of referring to non-volitional processes, i.e., processes not controlled by the subject. Two subclasses can be distinguished. The first class contains verbs of involuntary reaction either not involving motion (e.g., *schaudern* "shudder", *zittern* "jitter", and *beben* "tremble"), or involving motion (e.g., *torkeln* "totter", *taumeln* "stagger", or *wackeln*, "waggle"). Both types of verbs select *haben*:

(3.12) a. Die Frau *ist/hat angstvoll gezittert.
the woman is/has fearfully jittered
"The woman jittered with fear."
b. Die Frau ?ist/hat etwas getorkelt.
the woman is/has a-bit tottered
"The woman tottered a bit."

systematically alternate with reflexive assume position verbs, e.g., *sich setzen* "sit", *sich stellen* "stand", *sich hocken* "squat", or *sich knieen* "kneel". The contrast is illustrated by the following example:

(3.9) a. *Das Kind ist/hat auf den Boden gehockt.
the child is/has on the floor squated
"The child squatted on the floor."
b. Das Kind *ist/hat sich auf den Boden gehockt.
the child is/has self on the floor squated
"The child squatted on the floor."

Note that reflexive verbs unambiguously select *haben* in German.

Verbs of involuntary reaction involving motion can be telicized by adding a directional adverbial. They then behave like motion verbs and selection *sein*:[3]

(3.13) a. Die Frau *ist/hat in der Wohnung getorkelt.
the woman is/has in the flat tottered
"The woman tottered in the flat."

b. Die Frau ist/*hat in die Wohnnung getorkelt.
the woman is/has into the flat tottered
"The woman tottered into the flat."

The second class of uncontrolled process verbs includes verbs of emission such as *rumpeln* "rumble", *brummen* "buzz", and *klappern* "rattle". These verbs typically select *haben*:

(3.14) Der Zug *ist/hat laut gerumpelt.
the train is/has noisily rattled
"The train rattled noisily."

Again, verbs of this type can be telicized by adding a directional adverbial like *in den Bahnhof* "into the station" (see (3.15a)). In this case, the verb is interpreted as a motion verb (where the motion includes a sound emission), and we find a *sein* preference. In the presence of a positional adverbial such as *im Bahnhof* "in the station", we get an atelic interpretation and a *haben* preference (see (3.15b)). (This phenomenon is documented in Levin and Rappaport Hovav 1995.)

(3.15) a. Der Zug ist/*hat in den Bahnhof gerumpelt.
the train is/has in the station rattled
"The train rattled into the station."

b. Der Zug *ist/hat im Bahnhof gerumpelt.
the train is/has into the station rattled
"The train rattled in the station."

The fact that these auxiliary shifts occur indicates that uncontrolled process verbs are peripheral unergatives.

Controlled Process (Motional) Verbs in this class describe the physical motion of the subject and usually denote a manner of motion. Motion verbs are generally unergative in French, Italian, and Dutch, and select HAVE. In German, however, motion verbs tend to select BE:

(3.16) Die Frau ist/?hat schnell geschwommen.
the man is/has rapidly swam
"The woman swam rapidly."

[3] In Section 3.3.1 we will argue (on the basis of the outcome of Experiment 2) that verbs like *taumeln* 'totter' do not in fact belong to the class of uncontrolled process verbs. Rather, they are verbs of manner of motion, which explains why they display the alternations typical for controlled process (motional) verbs. This includes telicization by a directional PP as illustrated in (3.13).

This seems to be a gradient phenomenon, i.e., some motion verbs allow *haben* to a certain degree:

(3.17) a. Die Nachbarin ist/?hat langsam geschlurft.
the neighbor is/has slowly scuffled
"The neighbor scuffled slowly."

b. Die Tänzerin ist/?hat langsam getanzt.
the dancer is/has slowly danced
"The dancer danced slowly."

This indicates that motion verbs are peripheral unaccusatives. This is also supported by the fact that they undergo auxiliary shifts, consider the contrasts (3.18):

(3.18) a. Die Frau ist/*hat ans Ufer geschwommen.
the woman is/has to the shore swam
"The woman swam to the shore."

b. Die Frau *ist/hat im Fluss geschwommen.
the woman is/has in the river swam
"The woman swam in the river."

For controlled process (motional) verbs such as *schwimmen* "swim" in (3.18), a telic reading induces an auxiliary preference for *sein*, while an atelic reading induces a preference for *haben*. The telic reading can be triggered by a directional adverbial such as *ans Ufer* "to the shore" (see (3.18a)), while an atelic reading can be triggered by a positional adverbial such as *im Fluss* "in the river" (see (3.18b)). (This phenomenon is documented in Levin and Rappaport Hovav 1995.)

Another indicator of the peripheral status of these verbs is the fact that they are subject to dialectal variation (discussed in Section 3.2.1.3).

Controlled Process (Non-Motional) Verbs in this class denote non-motional, agentive processes. Sorace (2000) classifies them as core unergatives as they are consistently unergative across languages. Examples include *reden* "talk", *warten* "wait", *telefonieren* "phone", or *arbeiten* "work". These verbs select *haben* in German:

(3.19) Die Lehrerin *ist/hat dauernd geredet.
the teacher is/has continuously talked
"The teacher talked continuously."

No gradience is attested for the auxiliary selection behavior of the verbs in this class; the alternative auxiliary *sein* is seriously unacceptable.

3.2.1.2. Impersonal Passives

A number of unergative diagnostics other than auxiliary selection have been proposed for German (Grewendorf 1989; Seibert 1993). In this thesis we focus on impersonal passive formation,

which has been claimed to be possible with unergative verbs, but not with unaccusative ones (Levin and Rappaport Hovav 1995; Zaenen 1993). Examples for the verb classes discussed in the previous section are given in (3.20) and (3.21). Core unaccusatives like change of location verbs disallow impersonal passives; in peripheral unaccusatives like continuation of state and existence of state verbs, the acceptability of impersonal passives is reduced:

(3.20) a. Change of Location
*Es wurde schnell entkommen.
it was quickly escaped
b. Change of State
*Es wurde langsam errötet.
it was slowly blushed
c. Continuation of State
?Es wurde kurz verweilt.
it was briefly stayed
d. Existence of State (Positional)
?Es wurde würdevoll gekniet.
it was dignified kneeled

For core unergatives like controlled process (non-motional) verbs, impersonal passives are fully acceptable. The judgments for the other unergative classes vary:

(3.21) a. Uncontrolled Process (Involuntary Reaction)
*Es wurde angstvoll gezittert.
it was fearfully jittered
b. Uncontrolled Process (Emission)
?Es wurde laut gerumpelt.
it was noisily rattled
c. Controlled Process (Motional)
?Es wurde schnell geschwommen.
it was rapidly swam
d. Controlled Process (Non-Motional)
Es wurde dauernd geredet.
it was continuously talked

3.2.1.3. Dialect Variation

Sorace's (2000) account predicts crosslinguistic variation in the unergative/unaccusative behavior of peripheral, but not of core verbs. Under the assumption that dialect variation is an instance of crosslinguistic variation, we would expect the auxiliary selection behavior of peripheral verbs to be subject to dialectal differences, while the auxiliary selection behavior of core verbs should be stable across dialects. There is evidence for this hypothesis from dialects of Italian, such as the Friul and Veneto dialects discussed by Haider and Rindler-Schjerve

(1987), or Sardinian discussed by Sorace (2000). Further evidence on dialect variation comes from the study of auxiliary selection in Paduan reported by Sorace and Cennamo (2000).

Data from dialects of German is only mentioned in passing by Haider and Rindler-Schjerve (1987), who observe that the verbs *sitzen* "sit", *liegen* "lie", and *stehen* "stand" select *haben* in northern varieties of German, while they select *sein* in southern varieties (Bavarian and Austrian dialects). All three verbs are existence of state verbs, i.e., they are peripheral verbs for which crosslinguistic variation is expected under Sorace's (2000) account. (Note that most existence of state verbs select *hebben* "have" also in Dutch.)

Dialect differences have also been observed for other peripheral classes, such as the controlled process (motional) class (Grewendorf 1989: 10): verbs like *schwimmen* "swim", *wandern* "hike", or *rennen* "run" seem to select *haben* in southern dialects, while they prefer *sein* in northern dialects.

3.2.2. Introduction

The present experiment elicits judgments for auxiliary selection and impersonal passive formation in German. The experimental design is based on the Auxiliary Selection Hierarchy described in the previous section (see (3.1)). The aim is to test Sorace's (2000) claim that core unaccusative/unergative verbs exhibit binary auxiliary selection behavior, while peripheral verbs show gradient auxiliary selection preferences. We will elicit data from speakers of two dialectal variants of German, which enables us to test the additional claim that peripheral, but not core, verbs are subject to crosslinguistic (here, crossdialectal) differences.

In Experiments 2 and 3 we will refine the semantic classification used in the present experiment, and also test for animacy effects and telicity effects induced by prefixes and adverbial modifiers.

3.2.3. Predictions

3.2.3.1. Constraints

We predict that auxiliary selection in German is sensitive to the unaccusative/unergative distinction. More precisely, we expect that the semantic class a verb belongs to has an influence on its auxiliary selection behavior, i.e., we predict a significant interaction of verb class and auxiliary.

According to Sorace's (2000) account, core unaccusatives (such as change of location or change of state verbs) select the auxiliary *sein* "be", while core unergatives (such as controlled process (non-motional) verbs) select *haben* "have". Binary auxiliaries selection preferences are expected for core verbs, i.e., the "right" auxiliary should be fully acceptable, while the "wrong" one should lead to strong unacceptability. Peripheral verbs, on the other hand, are predicted to be less stable in their auxiliary selection behavior. These verbs should exhibit

gradient selection preferences, i.e., they should also allow the "wrong" auxiliary to a certain degree. This prediction can be tested by carrying out post-hoc analysis on the interaction of verb class and auxiliary.

Impersonal passive formation is another phenomenon that is sensitive to the unaccusative/unergative distinction. We predict a significant main effect of verb class for impersonal passives. As impersonal passive formation is a less reliable diagnostic of unergativity, we do not expect a perfect match between the acceptability of impersonal passives and auxiliary selection preferences.

3.2.3.2. Constraint Types

Furthermore, Sorace (2000) predicts crosslinguistic differences in the auxiliary selection behavior of peripheral verbs, but not of core verbs. This prediction can be tested in the present experiment by comparing speakers of different dialects of German. We expect core verbs to be stable across dialects, while peripheral verbs should exhibit dialectal variation. In particular, we predict dialect differences for the classes existence of state and controlled process (motional), in line with the relevant observations in the literature (see Section 3.2.1.3). This prediction will be tested using planned comparisons on the auxiliary preferences of these two classes.

3.2.4. Method

3.2.4.1. Subjects

Twenty-three native speakers of German participated in the experiment. The subjects were recruited over the Internet by postings to relevant newsgroups and mailing lists. Participation was voluntary and unpaid. Subjects had to be linguistically naive, i.e., neither linguists nor students of linguistics were allowed to participate.

The data of one subject were excluded because she was bilingual (by self-assessment). The data of another subject were excluded because she was a linguist (by self-assessment). The data of a third subject were eliminated after an inspection of the responses showed that she had not completed the task adequately.[4]

This left 20 subjects for analysis. Of these, 15 subjects were male, five female; two subjects were left-handed, 18 right-handed. The age of the subjects ranged from 19 to 45 years, the mean was 29.7 years.

[4] In all experiments reported in Chapters 3–5, subjects were excluded based on response times and response ranges. Chapter 5 contains a more detailed description of the data recorded by the experimental software for this purpose.

3.2.4.2. Materials

Training Materials The experiment included a set of training materials that were designed to familiarize subjects with the magnitude estimation task. The training set contained six horizontal lines. The range of largest to smallest item was 1:10. The items were distributed evenly over this range, with the largest item covering the maximal window width of the web browser. A modulus item in the middle of the range was provided.

Practice Materials A set of practice items was used to familiarize subjects with applying magnitude estimation to linguistic stimuli. The practice set consisted of six sentences that were representative of the test materials. A wide spectrum of acceptability was covered, ranging from fully acceptable to severely unacceptable. A modulus item in the middle of the range was provided.

Test Materials The experiment included two subdesigns. The first subdesign tested auxiliary preferences and crossed the factors verb class (*Verb*) and auxiliary (*Aux*). The factor *Verb* included eight levels, corresponding to the verb classes listed in Table 3.2. The factor *Aux* had two levels, *sein* and *haben*. This yielded a total of $Verb \times Aux = 8 \times 2 = 16$ cells. Eight lexicalizations were constructed for each cell, involving the verbs given in Table 3.2, an animate subject, and an adverb of manner (see (3.2)–(3.19)).[5] This yielded a total of 128 stimuli.

The second subdesign tested the acceptability of impersonal passives, with verb class as the only factor. This factor had the same levels as in the first subexperiment. This time, however, the verbs embedded in an impersonal passive construction (see (3.20) and (3.21)). The same eight lexicalizations as in the first subexperiment were used for each class, creating a total of 64 stimuli.

A set of 16 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

To control for possible effects from lexical frequency, the verb classes were matched for frequency. Verb frequencies were obtained from a lemmatized version of the Frankfurter Rundschau corpus (40 million words of newspaper text) and the average verb frequency for each verb class was computed. An ANOVA confirmed that these average frequencies were not significantly different from each other.

3.2.4.3. Procedure

The method used was magnitude estimation as proposed by Stevens (1975) for psychophysics and extended to linguistic stimuli by Bard et al. (1996). Each subject took part in an experimental session that lasted approximately 15 minutes and consisted of a training phase, a practice

[5] An exception is the uncontrolled process (emission) class, as the verbs in this class do not allow animate subjects.

Table 3.2: Verb classes and class members (Experiment 1)

unaccusative	
change of location	aufsteigen “climb”, entkommen “escape”, zurückkommen “come back”, ankommen “arrive”, abreisen “depart”, flüchten “flee”, weggehen “go away”, vorrücken “move forward”
change of state	erscheinen “appear”, erblassen “become pale”, nervös werden “become nervous”, versterben “die”, erröten “blush”, erkalten “become cold”, wachsen “grow”, verschwinden “disappear”
continuation of state	dahinvegetieren “vegetate”, überdauern “outlast”, aushalten “endure”, weiterexistieren “continue exiting”, weiterleben “continue living”, überleben “survive”, verharren “persist”, verweilen “stay”
existence of state (positional)	herumstehen “stand about”, herumhängen “hang about”, knien “kneel”, kauern “crouch”, baumeln “dangle”, schweben “hover”, sitzen “sit”, hocken “squat”
unergative	
uncontrolled process (involuntary reaction)	torkeln “totter”, taumeln “stagger”, wackeln “waggle”, schwanken “wobble”, schaudern “shudder”, beben “tremble”, zittern “jitter”, schlottern “shiver”
uncontrolled process (emission)	rumpeln “rumble”, klappern “rattle”, brummen “buzz”, quietschen “squeak”, rattern “clatter”, tuckern “tap”, surren “whir”, ächzen “moan”
controlled process (motional)	schwimmen “swim”, wandern “hike”, schlurfen “shuffle”, rennen “run”, tanzen “dance”, klettern “climb”, kriechen “creep”, hüpfen “bounce”
controlled process (non-motional)	reden “talk”, dozieren “lecture”, plaudern “chat”, warten “wait”, arbeiten “work”, telefonieren “telephone”, nachgeben “give in”, mitspielen “play”,

phase, and an experimental phase. The experiment was self-paced, though response times were recorded to allow the data to be screened for anomalies.

The experiment was conducted remotely over the Internet. The subject accessed the experiment using his or her web browser. The browser established an Internet connection to the experimental server, which was running WebExp 2.1 (Keller, Corley, Corley, Konieczny, and Todirascu 1998), an interactive software package for administering web-based psychological experiments. (The reliability and validity of web-based experimentation is assessed in Chapter 5. This chapter also contains a detailed description of the experimental software.)

Instructions Before the actual experiment started, a set of instructions in German was presented. The instructions first explained the concept of numeric magnitude estimation of line length. Subjects were instructed to make estimates of line length relative to the first line they would see, the reference line. Subjects were told to give the reference line an arbitrary number, and then assign a number to each following line so that it represented how long the line was in proportion to the reference line. Several example lines and corresponding numeric estimates were provided to illustrate the concept of proportionality.

Then subjects were told that linguistic acceptability could be judged in the same way as line length. The concept of linguistic acceptability was not defined; instead, examples of

acceptable and unacceptable sentences were provided, together with examples of numeric estimates.

Subjects were told that they could use any range of positive numbers for their judgments, including decimals. It was stressed that there was no upper or lower limit to the numbers that could be used (exceptions being zero or negative numbers). Subjects were urged to use a wide range of numbers and to distinguish as many degrees of acceptability as possible. It was also emphasized that there were no "correct" answers, and that subjects should base their judgments on first impressions, not spending too much time to think about any one sentence. The full set of instructions is listed in Appendix A.

Demographic Questionnaire After the instructions, a short demographic questionnaire was administered. The questionnaire included name, email address, age, sex, handedness, academic subject or occupation, and language region. Handedness was defined as "the hand you prefer to use for writing", while language region was defined as "the place (town, federal state, country) where you learned your first language". The results of the questionnaire were reported in Section 3.2.4.1.

Training Phase The training phase was meant to familiarize subjects with the concept of numeric magnitude estimation using line lengths. Items were presented as horizontal lines, centered in the window of the subject's web browsers. After viewing an item, the subject had to provide a numeric judgment over the computer keyboard. After pressing Return, the current item disappeared and the next item was displayed. There was no possibility to revisit previous items or change responses once Return had been pressed. No time limit was set for either the item presentation or for the response.

Subjects first judged the modulus item, and then all the items in the training set. The modulus was the same for all subjects, and it remained on the screen all the time to facilitate comparison. Items were presented in random order, with a new randomization being generated for each subject.

Practice Phase This phase allowed subjects to practice magnitude estimation of linguistic acceptability. Presentation and response procedure was the same in the training phase, with linguistic stimuli being displayed instead of lines. Each subject judged the whole set of practice items.

As in the training phase, subjects first judged the modulus item, and then all the items in the practice set. The modulus was the same for all subjects, and it remained on the screen all the time to facilitate comparison. Items were presented in random order, with a new randomization being generated for each subject.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in the practice phase.

Eight test sets were used: each test set contained one lexicalization for each of the 16 cells in the first subdesign, and one lexicalization for each of the eight cells in the second subdesign, i.e., a total of 24 items. Lexicalizations were assigned to test sets using a Latin square covering the full set of items.[6]

As in the practice phase, subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 40 test items: 24 experimental items and 16 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to one of the test sets.

3.2.5. Results

The data were normalized by dividing each numeric judgment by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. Then the data were transformed by taking the decadic logarithm. This transformation ensures that the judgments are normally distributed and is standard practice for magnitude estimation data (Bard et al. 1996; Lodge 1981). All analyses were conducted on the normalized, log-transformed judgments.

All the figures in Chapters 3–5 display means of normalized, log-transformed judgments, together with standard errors. Appendix C contains the descriptive statistics for all experimental results.

3.2.5.1. Constraints

Auxiliary Selection The mean judgments for each verb class for both auxiliaries are graphed in Figure 3.1. An ANOVA revealed a main effect of *Aux* (auxiliary), which however was significant only by subjects ($F_1(1,19) = 15.939,\ p = .001$; $F_2(1,7) = 2.210,\ p = .181$). The main effect of *Verb* (verb class) was not significant. As predicted, a highly significant interaction of *Aux* and *Verb* was obtained ($F_1(7,133) = 22.867,\ p < .0005$; $F_2(7,49) = 18.822,\ p < .0005$).

[6] In a Latin square, the first cell is assigned to the first stimulus set using the first lexicalization, to the second stimulus set using the second lexicalization, etc., rotating through the complete set of materials. This is illustrated for an example with four lexicalization in (3.22), where S_i are stimulus sets, C_j are cells in the design, and L_k are lexicalizations.

(3.22)

	S_1	S_2	S_3	S_4
C_1	L_1	L_2	L_3	L_4
C_2	L_2	L_3	L_4	L_1
C_3	L_3	L_4	L_1	L_2
C_4	L_4	L_1	L_2	L_3
C_5	L_1	L_2	L_3	L_4
C_6	L_2	L_3	L_4	L_1
C_7	L_3	L_4	L_1	L_2
C_8	L_4	L_1	L_2	L_3

In cases where the number of the cells is greater than the number of lexicalization, the Latin square is simply repeated *n* times, provided that the number of cells is the *n*-th multiple of the number of lexicalizations. This is illustrated in the lower half of (3.22).

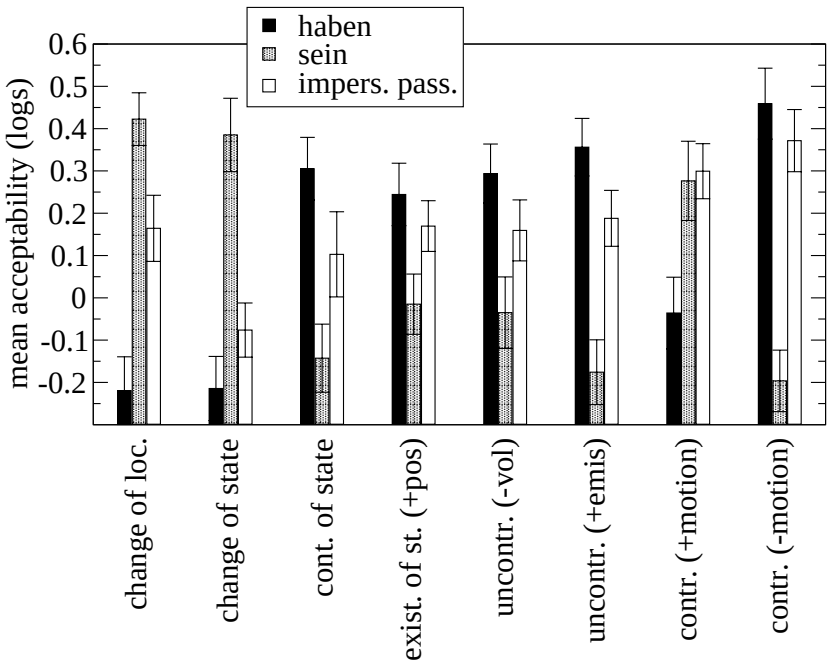

Figure 3.1: Mean judgments for auxiliary selection and impersonal passive (Experiment 1)

This confirms the hypothesis that auxiliary selection in German depends on the semantic class of the verb.

To further investigate the *Aux*/*Verb* interaction, a post-hoc Tukey test was conducted. The results of the Tukey test show which verb classes differed in auxiliary selection behavior. For *haben*, we found significant differences between the change of location class and the classes continuation of state ($\alpha < .01$), existence of state (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), controlled process (non-motional) ($\alpha < .01$), uncontrolled process (involuntary reaction) ($\alpha < .01$), and uncontrolled process (emission) ($\alpha < .01$). We also found significant differences between the change of state class and the classes continuation of state ($\alpha < .01$), existence of state (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), controlled process (non-motional) ($\alpha < .01$), uncontrolled process (involuntary reaction) (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), and uncontrolled process (emission) ($\alpha < .01$). A significant difference was also obtained between the controlled process (non-motional) and the controlled process (motional) class (by subjects, $\alpha < .01$ and by items, $\alpha < .05$) and between the controlled process (motional) and the uncontrolled process (emission) class (by subjects only, $\alpha < .05$).

For *sein*, there was a difference between the change of location class and the classes continuation of state ($\alpha < .01$), existence of state (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), controlled process (non-motional) ($\alpha < .01$), uncontrolled process (involuntary reaction) (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), and uncontrolled process (emission) ($\alpha < .01$). We also found significant differences between the change of state class and the classes contin-

uation of state (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), existence of state (by subjects only, $\alpha < .05$), controlled process (non-motional) ($\alpha < .01$), uncontrolled process (involuntary reaction) (by subjects only, $\alpha < .05$), and uncontrolled process (emission) ($\alpha < .01$). The difference between the continuation of state and the controlled process (motional) class was also significant ($\alpha < .05$). A significant difference was also obtained between the controlled process (motional) class and the controlled process (non-motional) class (by subjects, $\alpha < .01$ and by items, $\alpha < .05$) and the uncontrolled process (emission) class (by subjects, $\alpha < .01$ and by items, $\alpha < .05$).

Furthermore, the Tukey test shows which verb classes exhibit a significant difference between the acceptability of *haben* and *sein*. For the change of location class and the change of state class, *sein* was more acceptable than *haben* ($\alpha < .01$ in both cases). For the continuation of state class, *haben* was more acceptable than *sein* (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), while there was no significant difference between the auxiliaries for the existence of state class. *Haben* was more acceptable than *sein* for the controlled process (non-motional) and the uncontrolled process (emission) classes ($\alpha < .01$ in both cases), while there was no significant difference between the two auxiliaries for the controlled process (motional) and the uncontrolled process (involuntary reaction) classes.

Impersonal Passives The mean judgments for impersonal passives are also graphed in Figure 3.1. A separate ANOVA was conducted for the subexperiment on impersonal passives. A significant main effect of verb class was obtained ($F_1(7,133) = 5.068$, $p < .0005$; $F_2(7,49) = 4.265$, $p = .001$), which confirms our hypothesis that impersonal passives formation is sensitive to the semantic class of the verb.

A post-hoc Tukey test revealed that the acceptability of impersonal passives differed significantly for the change of state class and the controlled process (non-motional) class ($\alpha < .01$), the controlled process (motional) class (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), and the uncontrolled process (emission) class (by subjects only, $\alpha < .05$). The continuation of state and the controlled process (non-motional) class were also significantly different ($\alpha < .05$).

3.2.5.2. Constraint Types

To test the hypothesis that there is crosslinguistic variation in the auxiliary selection behavior of peripheral verbs, but not of core verbs, we divided the subjects into two dialect groups. As part of the personal details questionnaire, subjects had to specify a language region, i.e., the town, federal state, and country where they acquired their native language. Based on these answers we formed two groups: if the language region was in Austria, Switzerland or in a southern German federal state (Bavaria or Baden-Württemberg), then the subject was classified as a speaker of a southern dialect. All other subjects were classified as speakers of northern dialects. (No subjects stated language regions outside Austria, Switzerland, or Germany.) Ten subjects were speakers

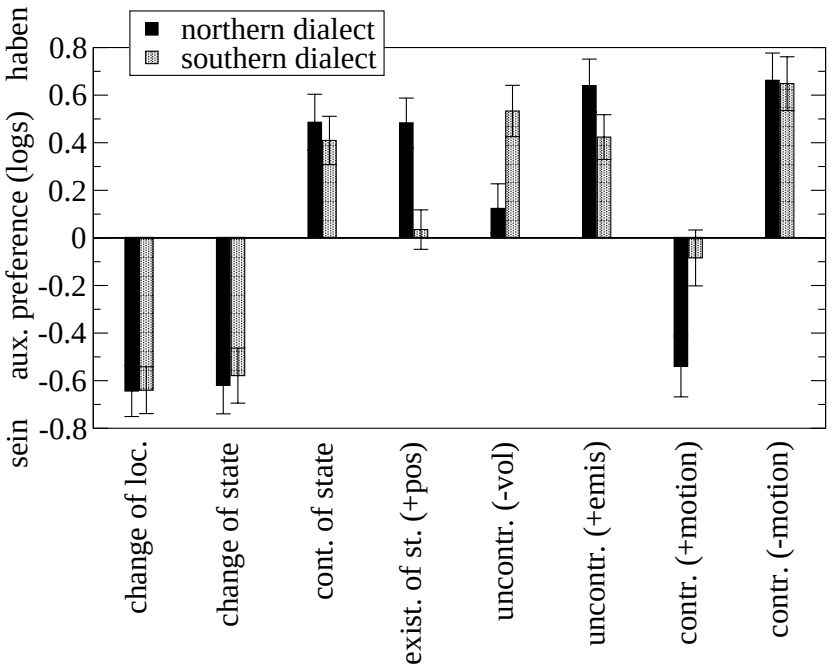

Figure 3.2: Mean judgments for auxiliary selection by dialect (Experiment 1)

of southern dialects, the other ten were speakers of northern dialects.

The auxiliary preferences for each verb class for both dialect groups are graphed in Figure 3.2. Note that this figure does not display absolute auxiliary selection judgments, but auxiliary preferences, i.e., the difference of the *sein* judgments and the *haben* judgments, rather than the absolute judgments.

An ANOVA on the auxiliary selection judgments used dialect as a between-subjects variable.[7] (Only a by-subject analysis could be conducted because the by-dialect split resulted in empty cells, i.e., there were some lexicalizations that were not represented in both dialect groups.) This ANOVA yielded a significant main effect of *Aux* ($F_1(1,18) = 15.269$, $p = .001$). There were no main effects of verb class or dialect. The interaction of verb class and auxiliary was significant ($F_1(7,126) = 24.057$, $p < .0005$), as was the interaction of verb class and dialect ($F_1(7,126) = 2.609$, $p = .015$). We also found a marginal three way interaction of verb class, auxiliary selection, and dialect ($F_1(7,126) = 11.989$, $p = .062$). There was no interaction of *Aux* and dialect.

We carried out planned comparisons on the classes for which we predicted a dialect effect, i.e., the existence of state and controlled process (motional) verbs.[8] As two planned

[7] This ANOVA replicates the ANOVA in Section 3.2.5.1, but includes dialect as an additional factor. This explains why the degrees of freedom and the F-values differ slightly from the ones in Section 3.2.5.1. Note that this is not a case of multiple tests on the same data (rather we refine an existing test), hence there is no need to adjust the p-value.

[8] Planned comparisons instead of post-hoc tests were used as we had a clear prediction regarding which

comparisons were carried out, we adjusted the *p*-value according to the Bonferroni method, i.e., we assumed $p = .025$ as our significance level. For both classes, we found a marginally significant interaction of dialect and auxiliary ($F_1(1,18) = 4.274$, $p = .053$ and $F_1(1,18) = 4.145$, $p = .057$, respectively).

Furthermore, we tested if dialect has an influence on impersonal passive formation. This ANOVA yielded a significant main effect of verb class ($F_1(7,126) = 5.234$, $p < .0005$), but the main effect of dialect and the interaction of dialect and verb class were not significant.

3.2.6. Discussion

3.2.6.1. Constraints

We demonstrated that the semantic class a verb belongs to has an influence on auxiliary selection and impersonal passive formation in German. Unaccusative verbs were shown to generally prefer the auxiliary *sein* "be", while unergative verbs generally prefer *haben* "have". We also found that impersonal passives were more acceptable with unergative verbs than with unaccusative verbs.

3.2.6.2. Constraint Types

Following Sorace (2000), we distinguished two types of verbs: core verbs and peripheral verbs. As predicted, peripheral verbs exhibited gradient auxiliary selection preferences and were subject to crosslinguistic variation. Core verbs, on the other hand, showed a binary preference for one auxiliary that was crosslinguistically stable.

In line with Sorace's (2000) predictions, we found that change of state verbs and change of location verbs were core unaccusatives, while controlled process (non-motional) verbs were core unergatives. Examples for peripheral unaccusatives are the verbs in the existence of state class. There is an overall preference for *haben* in this class (see Figure 3.1), which however is subject to dialect variation (see Figure 3.2): speakers of northern dialects prefer *haben*, while speakers of southern dialects have no clear preference for either *sein* or *haben*. Another interesting case is the continuation of state class, which exhibits a preference for *haben* in both dialects. Verbs of this type, however, prefer *sein* in other Germanic languages (e.g., in Dutch, see Sorace and Vonk 1998).

As for peripheral unergative verbs, controlled process (motional) verbs show an overall weak preference for *sein*, while uncontrolled process (involuntary reaction) verbs show an overall weak preference for haben (see Figure 3.1). The fact that the auxiliary selection preferences are rather weak is in line with the peripheral status of these verbs, as is the fact that

verb classes should show dialect effects (based on the theoretical literature, see Section 3.2.1.3). The planned comparisons have the advantage of being more selective (and hence more powerful) than a blanket Tukey test on the interaction of verb class, auxiliary selection, and dialect.

there is dialectal variation: *sein* is more acceptable for peripheral unergatives for speakers of northern dialects, while *haben* is judged more acceptable by speakers of southern dialects (see Figure 3.2). The class uncontrolled process (emission) does not fit into this pattern; it exhibits a clear *haben* preference, which is subject to only small dialectal differences.

Another prediction was that impersonal passive formation correlates with unergative/unaccusative status (Levin and Rappaport Hovav 1995; Zaenen 1993). This prediction was borne out: impersonal passives are significantly more acceptable for unergative verbs than for unaccusative verbs, which is in line with the relevant observations in the literature (Grewendorf 1989; Seibert 1993). However, there is considerable variation in the acceptability of impersonal passive formation across classes (see Figure 3.1). Also, we failed to find dialectal differences for impersonal passives. Both facts are in line with the claim that impersonal passive formation is a less reliable diagnostic of unergativity than auxiliary selection (Sorace 2000).

3.2.7. Conclusions

The present experiment investigated unaccusative/unergative verbs with respect to auxiliary selection and impersonal passive formation. We provided evidence for a subdivision into core and peripheral verbs, as hypothesized by Sorace (2000). Core verbs show a clear preference for one auxiliary and are immune to dialectal variation. Peripheral verbs exhibit gradient auxiliary selection preferences, i.e., they allow both auxiliaries to a certain degree. Also, we found that the auxiliary selection preferences of peripheral verbs are subject to dialect variation.

The results of this experiment give us to two empirical criteria for distinguishing core and peripherals verbs, in line with Sorace's (2000) predictions: gradient acceptability and crosslinguistic (here, crossdialectal) variation.

3.3. Experiment 2: Effect of Animacy and Telicity on Unaccusativity and Unergativity

Experiment 1 provided evidence for the distinction between core and peripheral verbs, based on the Auxiliary Selection Hierarchy (see (3.1)). It was demonstrated that peripheral verbs show gradience and crosslinguistic variation in their auxiliary selection behavior, while core verbs exhibit behavior that is binary and crosslinguistically stable.

The present experiment is designed to investigate two further influences on auxiliary selection, viz., telicity and animacy. In the light of the theoretical literature (e.g., Levin and Rappaport Hovav 1995; Sorace 2000), we expect peripheral verbs, but not core verbs, to be subject to telicity and animacy effects. In order to test this prediction, we will refine the classification of unaccusative and unergative verbs used in Experiment 1. In particular, we will investigate animacy effects that have been reported for certain verb classes, and telicity effects

that can be attributed to verb prefixes.

3.3.1. Background

Change of State Change of state verbs showed a clear preference for *sein* in Experiment 1. Previously, however, these verbs have been classified as peripheral verbs, which leads us to predict gradient auxiliary selection behavior. The failure to find gradience might be due to the fact that the change of state verbs included in Experiment 1 are mainly verbs that denote a change with a definite endpoint, such as *erscheinen* "appear" or *erblassen* "become pale". Only a few verbs that refer to an incremental change (such as *wachsen* "grow") were part of the stimuli (see also Table 3.2). To test this hypothesis, we used a different set of verbs for the change of state class in the present experiment. We included only verbs that clearly denote to an incremental change, such as *rosten* "rust" or *blühen* "blossom".

Note that some change of state verbs allow prefixing, which intuitively changes their auxiliary selection behavior. The prefix seems to give the verb a telic reading that implies an endpoint for the change of state denoted by the verb. As examples consider (3.23) and (3.24). In the unprefixed (a) variants, the verb has an atelic incremental change reading, while the prefix in the (b) variant induces an telic reading that implies a definite endpoint of the change.

(3.23) a. Die Dose ?ist/hat sofort gerostet.
the can is/has immediately rusted
"The can was rusting immediately."

b. Die Dose ist/*hat sofort verrostet.
the can is/has immediately rusted
"The can got rusty immediately."

(3.24) a. Die Rose ?ist/hat sofort geblüht.
the rose is/has immediately blossomed
"The rose was blossoming immediately."

b. Die Rose ist/*hat sofort erblüht.
the rose is/has immediately blossomed
"The rose blossomed immediately."

The (a) verbs prefer *haben*, but also allow *sein* to a certain degree, while the (b) verbs only allow *sein*. To verify this intuition, the present experiment included a set of change of state verbs that can occur either in a prefixed or in a non-prefixed from, corresponding to the (a) and (b) examples in (3.23) and (3.24) (see Table 3.3 for details).

Continuation of State Continuation of state verbs showed a clear preference for *haben* in Experiment 1. However, there is some evidence in the literature that animacy can have an effect on the auxiliary selection preference of continuation of state verbs, as shown by Sorace (2000) for Italian.

Intuitively, an animacy effect seems to exist also for continuation of state verbs in German, consider the examples in (3.25):

(3.25) a. Der Wanderer ?ist/hat kurz verweilt.
the hiker is/has briefly stayed
"The hiker stayed briefly."

b. Der Regen *ist/hat kurz angedauert.
the rain is/has briefly lasted
"The lasted briefly."

For animate subjects as in (3.25a), we find a preference for *haben*, but *sein* seems to be not completely unacceptable. For inanimate subjects such as in (3.25b), there seems to be a clear preference for *haben* and clear dispreference for *sein.*

We tested this intuition by including a set of continuation of state verbs with inanimate subjects in the present experiment, which allows comparison with preferences obtained in Experiment 1 for continuation of state verbs with animate subjects (see Table 3.3 for details).

Existence of State Experiment 1 dealt with positional verbs, a subclass of existence of state verbs. We found evidence for crossdialectal variation of in the auxiliary selection behavior of these verbs; speakers of northern dialects prefer *haben* with positional verbs, while speakers of southern dialects allow both auxiliaries. The fact that we find dialectal variation for positional verbs confirms that these verbs are peripheral unaccusatives.

Another argument for the peripheral status of positional verbs is the fact that they are subject to animacy effects (see Sorace 2000 for Italian). With an animate subjects, these verbs allow a volitional reading that denotes the act of maintaining a position. Inanimate subjects, on the other hand, only allow a non-volitional reading, which simply denotes the position the subject is in.

It is possible that similar effects exist in German. As an example, consider (3.26), where the animate (a) example intuitively exhibits a slight preference for *haben*, while the inanimate (b) example exhibits a slight preference for *sein.*

(3.26) a. Die Täterin ?ist/hat betreten dagestanden.
the offender is/has sheepishly stood there
"The offender stood there sheepishly."

b. Der Korb ist/?hat unbeachtet dagestanden.
the basket is/has unnoticed stood there
"The basked stood there unnoticed."

The present experiment attempts to verify this observation by testing a set of positional verbs with both animate and inanimate subjects (see Table 3.3 for details).

Uncontrolled Process Verbs denoting uncontrolled, involuntary processes showed gradient auxiliary selection behavior in Experiment 1, with a weak preference for *haben.* However, there

seems to be an unexplained dialect difference for this verb class (see Figure 3.2); speakers of southern dialects preferred *haben*, while speakers of northern dialects allowed both *haben* and *sein*. (Note, however, that we did not test the significance of this effect. As it was unexpected, a planned comparisons could not be used.)

As detailed in Section 3.2.1.1, the uncontrolled process (involuntary reaction) class contains two types of verbs, viz., ones that denote a process involving motion (such as *torkeln* "totter" or *taumeln* "stagger"), and ones that do not involve motion (such as *schaudern* "shudder" or *zittern* "jitter"). Intuitively, the involuntary non-motion verbs show a preference for *haben*, while involuntary motion verbs allow both auxiliaries to a certain degree (see also (3.13)).

If we assume that involuntary motion verbs behave like verbs in the controlled process (motional) class, then we have an explanation for the dialect effect: in Experiment 1, we found that controlled process (motional) verbs prefer *sein* in southern dialects, but allow both *sein* and *haben* in southern dialects—it seems that a similar dialect effect was present in the uncontrolled (involuntary reaction) class. However, it was attenuated by the fact that non-motion verbs were also included in this class. The present experiment removes this confound by including a separate class with uncontrolled process, involuntary reaction, non-motional verbs (see Table 3.3 for details).

Change of Location and Controlled Process (Non-Motional) Change of location and controlled process (non-motional) verbs represent the core classes for unaccusative and unergative verbs, respectively, and are expected to show binary auxiliary selection preferences and no dialect variation. These two classes were included as a control condition in the present experiment; their auxiliary selection preference give us a standard against which to compare the auxiliary selection behavior of the other verb classes.

3.3.2. Introduction

The design of the present experiment is modeled on that of Experiment 1. We elicit judgments for auxiliary selection and impersonal passive formation in German, based on the refined classification described in the previous section. This refined classification allows us to test for telicity effects induced by prefixing in the change of state class. Furthermore, we will establish whether animacy has an effect on auxiliary selection for continuation of state and existence of state verbs. Animacy and telicity effects are associated with the peripheral status in the Auxiliary Selection Hierarchy of Sorace (2000), and hence will provide further evidence for the core/periphery distinction. Moreover, the present experiment includes uncontrolled, involuntary process verbs to eliminate a confound that was present in this class in Experiment 1. Change of location and controlled process, non-motional verbs will be included as controls, as they are core unaccusatives and core unergatives, respectively.

We will again elicit data from speakers of two dialectal variants of German, which will allow us to confirm he dialect effects found in Experiment 1.

3.3.3. Predictions

3.3.3.1. Constraints

For change of state verbs, we predict that adding a prefix will change the auxiliary selection preference form *haben* to *sein*, as the prefixed version of a change of state verb only allows a telic interpretation.

For continuation of state verbs, we predict that the use of inanimate subjects will change the auxiliary selection preference, in line with claims in the theoretical literature. Also for existence of state (positional) verbs, we expect the animacy of the subject to influence auxiliary selection preference, as only an animate subject allow a volitional (maintain position) reading.

Change of location and controlled process (non-motional) verbs were included as controls. These classes should show binary auxiliary selection behavior (in accordance with Experiment 1).

3.3.3.2. Constraint Types

The evidence from dialect variation is predicted to confirm the core/periphery classification. The core verbs should be immune to dialect effects. For peripheral verbs, on the other hand, we expect dialect effects similar to the ones found in Experiment 1, i.e., existence of state (positional) verbs should vary in auxiliary preference between speakers of northern and southern dialects of German. In addition, we expect dialect difference for change of state (no prefix) verbs, based on the observations in Section 3.3.1. These predictions will be tested using planned comparisons on the auxiliary preferences of these two classes.

3.3.4. Method

3.3.4.1. Subjects

Twenty-seven native Speakers of German from the same population as in Experiment 1 participated in the experiment. None of the subjects had previously participated in Experiment 1.

The data of two subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 25 subjects for analysis. Of these, 17 subjects were male, eight female; 22 subjects were right-handed, three left-handed. The age of the subjects ranged from 16 to 41 years, the mean was 27.3 years.

3.3.4.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials In analogy to Experiment 1, the present experiment included two subdesigns. The first subdesign tested auxiliary preferences and crossed the factors verb class (*Verb*) and auxiliary (*Aux*). The factor *Verb* included eight levels, corresponding to the verb classes listed in Table 3.3. The factor *Aux* had two levels, *sein* and *haben*. This yielded a total of $Verb \times Aux = 8 \times 2 = 16$ cells. Eight lexicalizations were constructed for each cell, involving the verbs given in Table 3.3 and an adverb of manner. Depending on the verb class, the subject was either animate or inanimate, as stated in Table 3.3. This yielded a total of 128 stimuli.

The second subdesign tested the acceptability of impersonal passives, with verb class as the only factor. This factor had the same levels as in the first subexperiment. This time, however, the verbs embedded in an impersonal passive construction (see (3.20) and (3.21)). The same eight lexicalizations as in the first subexperiment were used for each class, creating a total of 64 stimuli.

A set of 24 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

The verb classes were matched for frequency using the same procedure as in Experiment 1.

3.3.4.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions, Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

Eight test sets were used: each test set contained one lexicalization for each of the 16 cells in the first subdesign, and one lexicalization for each of the eight cells in the second subdesign, i.e., a total of 24 items. Lexicalizations were assigned to test sets using Latin squares. Three separate Latin squares were applied: one for the *haben* condition, one for the *sein* condition, and one for the impersonal passives.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 48 test items: 24 experimental items and 24 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to one of the test sets.

Table 3.3: Verb classes and class members (Experiment 2)

unaccusative	
change of location (animate)	aufsteigen “climb”, entkommen “escape”, zurückkommen “come back”, ankommen “arrive”, abreisen “depart”, flüchten “flee”, weggehen “go away”, vorrücken “move forward”
change of state (no prefix, inanimate)	rosten “rust”, modern “rot”, faulen “rot”, schimmeln “become mouldy”, welken “wilt”, blühen “bloom”, keimen “germinate”, wachsen “grow”, schwellen “swell”, sinken “sink”, steigen “rise”
change of state (prefix, inanimate)	verrosten “rust”, vermodern “rot”, verfaulen “rot”, verschimmeln “become mouldy”, verwelken “wilt”, verblühen “bloom”, aufkeimen “germinate”, anwachsen “grow”, versinken “sink”, anschwellen “swell”, ansteigen “rise”
continuation of state (inanimate)	dauern “last”, andauern “last”, fortdauern “last”, halten “last”, anhalten “continue”, reichen “suffice”, ausreichen “suffice”, genügen “suffice ”
existence of state (positional, animate)	stehen “stand”, dastehen “stand”, herumstehen “stand about”, herumhängen “hang about”, baumeln “dangle”, liegen “lie”, herumliegen “lie about”, daliegen “lie”, schweben “hover”
existence of state (positional, inanimate)	stehen “stand”, dastehen “stand”, herumstehen “stand about”, herumhängen “hang about”, baumeln “dangle”, liegen “lie”, herumliegen “lie about”, daliegen “lie”, schweben “hover”
unergative	
uncontrolled process (involuntary reaction, nonmotional, animate)	schaudern “shudder”, beben “tremble”, zittern “jitter”, schlottern “shiver”, zucken “convulse”, schwitzen “sweat”, gähnen “yawn”, keuchen “wheeze”, husten “cough”, niesen “sneeze”, schniefen “snuff”
controlled process (nonmotional, animate)	reden “talk”, dozieren “lecture”, plaudern “chat”, warten “wait”, arbeiten “work”, telefonieren “telephone”, nachgeben “give in”, mitspielen “play”,

3.3.5. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

3.3.5.1. Constraints

Auxiliary Selection The mean judgments for each verb class for both auxiliaries are graphed in Figure 3.3. An ANOVA revealed a significant main effect of *Aux* (auxiliary) ($F_1(1,24) = 25.327$, $p < .0005$; $F_2(1,7) = 16.372$, $p = .005$). The main effect of *Verb* (verb class) was significant by subjects only ($F_1(1,24) = 6.552$, $p < .0005$; $F_2(1,7) = 1.264$, $p = .228$). As predicted, a highly significant interaction of *Aux* and *Verb* was obtained ($F_1(7,168) = 43.684$, $p < .0005$; $F_2(7,49) = 34.757$, $p < .0005$).

To further investigate the *Aux*/*Verb* interaction, a post-hoc Tukey test was conducted. The results of the Tukey test show which verb classes differ in auxiliary selection behavior. For *haben*, we found significant differences between the change of location class and the change of state (no prefix) class (by subjects only, $\alpha < .01$), the continuation of state class

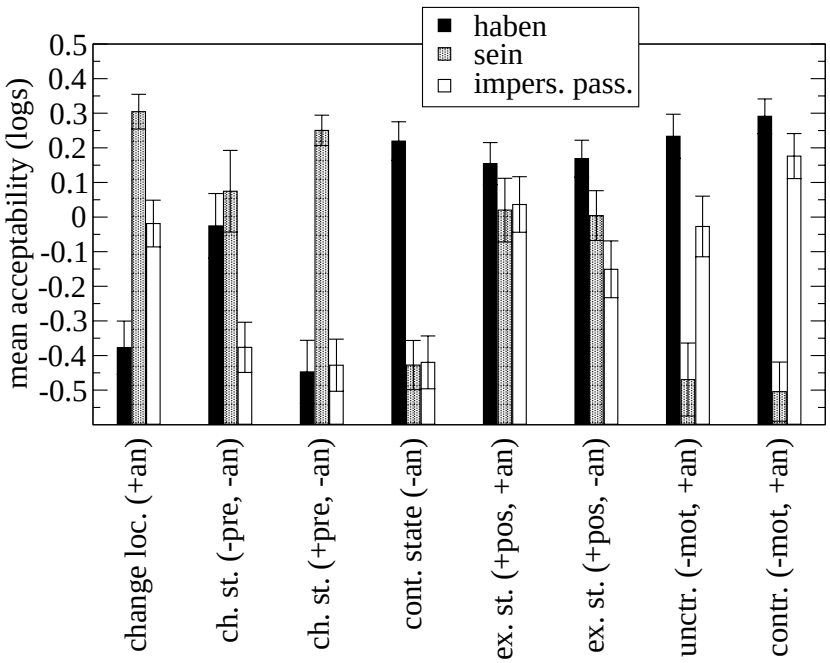

Figure 3.3: Mean judgments for auxiliary selection and impersonal passive (Experiment 2)

($\alpha < .01$), the existence of state (animate) class ($\alpha < .01$), the existence of state (inanimate) class ($\alpha < .01$), the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class ($\alpha < .01$). We also found significant differences between the change of state (prefix) class and the change of state (no prefix) class (by subjects, $\alpha < .01$ and by items, $\alpha < .05$), the continuation of state class ($\alpha < .01$), the existence of state (animate) class ($\alpha < .01$), the existence of state (inanimate) class ($\alpha < .01$), the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class ($\alpha < .01$).

For *sein*, there was a difference between the change of location class and the continuation of state class ($\alpha < .01$), the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class ($\alpha < .01$). We also found a difference between the change of state (no prefix) class and the continuation of state class ($\alpha < .01$), the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class ($\alpha < .01$). There was also a significant difference between the change of state (prefix) class and the continuation of state class ($\alpha < .01$), the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class ($\alpha < .01$). We also found a difference between the continuation of state class and the existence of state (animate) class ($\alpha < .01$), and the existence of state (inanimate) class ($\alpha < .01$). The difference between the existence of state (animate) class and the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class ($\alpha < .01$) was also significant, as was the difference between the existence of state (inanimate) class and the controlled process (non-

motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class ($\alpha < .01$).

Furthermore, the Tukey test shows which verb classes exhibit a significant difference between the acceptability of *haben* and *sein*. For the change of location class and the change of state (prefix) class, *sein* was more acceptable than *haben* ($\alpha < .01$ in both cases), while for the continuation of state, controlled process (motional), and uncontrolled process (involuntary reaction) classes, *haben* was more acceptable than *sein* ($\alpha < .01$ in all cases). For the remaining classes (change of state (no prefix), existence of state (animate), existence of state (inanimate)), no significant difference between the two auxiliaries was obtained.

Impersonal Passives The mean judgments for impersonal passives are also graphed in Figure 3.3. A separate ANOVA was conducted for the subexperiment on impersonal passives, yielding a significant main effect of verb class ($F_1(7,168) = 17.226$, $p < .0005$; $F_2(7,49) = 4.848$, $p < .0005$).

A post-hoc Tukey test revealed that the following classes are significantly different regarding the acceptability of impersonal passives: the change of location class and the change of state (no prefix) class (by subjects only, $\alpha < .01$), the change of state (prefix) class (by subjects only, $\alpha < .01$), and the continuation of state class (by subjects only, $\alpha < .01$). Furthermore, we found a difference between the change of state (no prefix) class and the existence of state (animate) class (by subjects only, $\alpha < .01$), the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class (by subjects only, $\alpha < .01$). There were also difference between the change of state (prefix) class and the existence of state (animate) class (by subjects, $\alpha < .01$, and by items, $\alpha < .05$), the existence of state (inanimate) class (by subjects only, $\alpha < .05$), the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class (by subjects only, $\alpha < .01$). The acceptability of impersonal passives differed for the continuation of state class and the existence of state (animate) class (by subjects only, $\alpha < .01$), the existence of state (inanimate) class (by subjects only, $\alpha < .05$), the controlled process (non-motional) class ($\alpha < .01$), and the uncontrolled process (involuntary reaction) class (by subjects only, $\alpha < .01$). Finally, the existence of state (inanimate) class and the controlled process (non-motional) class were also different (by subjects only, $\alpha < .01$).

3.3.5.2. Constraint Types

To test for dialect differences, we divided the subjects in speakers of southern and of northern dialects based on the same criteria as in Experiment 1. Thirteen subjects were speakers of southern dialects, twelve were speakers of northern dialects. The auxiliary preferences for each verb class for both dialect groups are graphed in Figure 3.4. Note that this figure displays auxiliary preferences, i.e., the difference of the *sein* judgments and the *haben* judgments, rather than the absolute judgments.

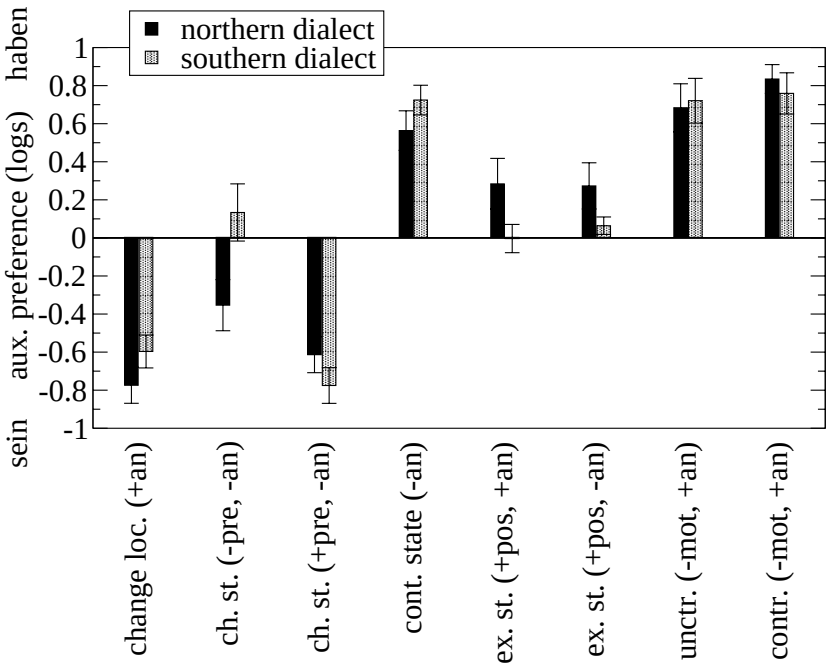

Figure 3.4: Mean judgments for auxiliary selection by dialect (Experiment 2)

We conducted an ANOVA on the auxiliary selection judgments with dialect as a between-subjects variable.[9] (Only a by-subject analysis could be conducted because the by-dialect split resulted in empty cells, i.e., there were some lexicalizations that were not represented in both dialect groups.) This ANOVA yielded a significant main effects of *Aux* ($F_1(1,23) = 24.384,\ p < .0005$) and of *Verb* ($F_1(1,23) = 6.781,\ p < .0005$). There was no main effect of dialect. The interaction of verb class and auxiliary also was significant ($F_1(7,161) = 45.720,\ p < .0005$), as was the three way interaction of *Aux*, *Verb*, and dialect ($F_1(7,161) = 2.118,\ p = .044$). We also found a marginal interaction of *Verb* and dialect ($F_1(7,161) = 1.838,\ p = .083$). There was no interaction of *Aux* and dialect.

We carried out planned comparisons on the classes for which we predicted a dialect effect, i.e., the change of state (no prefix) and existence of state (positional).[10] For the existence of state verbs, we combined the data for animate and inanimate subjects. As two planned comparisons were carried out, we adjusted the *p*-value according to the Bonferroni method, i.e., we assumed $p = .025$ as our significance level. For both classes, we found a marginally signif-

[9] This ANOVA replicates the ANOVA in Section 3.3.5.1, but includes dialect as an additional factor. This explains why the degrees of freedom and the *F*-values differ slightly from the ones in Section 3.3.5.1. Note that this is not a case of multiple tests on the same data (rather we refine an existing test), hence there is no need to adjust the *p*-value.

[10] Planned comparisons instead of post-hoc tests were used as we had a clear prediction regarding which verb classes should show dialect effects (based on the theoretical literature, see Section 3.2.1.3). The planned comparisons have the advantage of being more selective (and hence more powerful) than a blanket Tukey test on the interaction of verb class, auxiliary selection, and dialect.

icant interaction of dialect and auxiliary ($F_1(1,23) = 4.081$, $p = .055$ and $F_1(1,23) = 3.879$, $p = .061$, respectively).

Furthermore, we tested if dialect has an influence on impersonal passive formation. This ANOVA yielded a significant main effect of verb class ($F_1(7,161) = 16.998$, $p < .0005$), but the main effect of dialect and the interaction of dialect and verb class were not significant.

3.3.6. Discussion

3.3.6.1. Constraints

For change of state verbs, we predicted that adding a prefix would change the auxiliary selection preference form *haben* to *sein*. This prediction was borne out: we found that *haben* was significantly more acceptable for non-prefixed verbs than for prefixed verbs. The acceptability of *sein* was greater for prefixed verbs than for non-prefixed ones, although this difference failed to reach significance (see Figure 3.3).

For continuation of state verbs, we predicted that the use of inanimate subjects would change the auxiliary selection preference, in line with claims in the theoretical literature. This prediction was not borne out; as in Experiment 1, we found a clear *haben* preference for the continuation of state class. The only difference between animate (Experiment 1) and inanimate (Experiment 2) subjects with continuation of state verbs was that impersonal passives were less acceptable for inanimate subjects. This is not surprising, as the impersonal passive constructions requires an agentive interpretation, which is unavailable with verbs that prefer inanimate subjects (see Figures 3.1 and 3.3).

Also for existence of state (positional) verbs, we expected an effect of animacy on auxiliary selection preference, as only an animate subjects allow a volitional (maintain position) reading. Again, we failed to find this effect; the auxiliary selection preferences of existence of state verbs with animate and inanimate subjects were indistinguishable. Impersonal passives were again slightly less acceptable with inanimate subjects (see Figure 3.3).

Change of location and controlled process (non-motional) verbs were included as controls. For these classes, we found the same behavior as in Experiment 1: change of location verbs are core unaccusatives that strongly select for *sein*, while controlled process (non-motional) verbs are core unergative that have a clear *haben* preference.

3.3.6.2. Constraint Types

The present experiment showed dialect effects that are compatible with those reported in Experiment 1. For the existence of state class, we found that speakers of northern dialects prefer *haben*, while speakers of southern dialects allow both auxiliaries. The same pattern was observed in the present experiment, both for animate and inanimate subjects of existence of state verbs (see Figure 3.4). Furthermore, we found a dialect effect for the non-prefixed change of

state verbs. For these verbs, speakers of northern dialects prefer *sein*, while speakers of southern dialects prefer *haben*. It seems that verbs of this class receive a telic interpretation in northern dialects, but an atelic interpretation in southern dialects. (To our knowledge, the effect has not been documented in the literature so far.) For prefixed change of state verbs, on the other hand, no dialect effect was observed (see Figure 3.4). This points to the fact that the prefix induces a telic reading for these verbs and overrides the dialectal preference for a telic or atelic interpretation.

In this experiment, we had eliminated motion verbs from the class of uncontrolled process (involuntary reaction) verbs. The remaining verbs of involuntary reaction showed a clear preference for *haben* (see Figure 3.3). This confirms our assumption that only the motion verbs in this class allow *sein* as their auxiliary. A comparison of Figure 3.2 and Figure 3.4 shows that the dialect difference found in Experiment 1 for uncontrolled process (involuntary reaction) verbs disappeared in the present experiment. This is compatible with the assumption that only motion verbs (which were absent in the present experiment) exhibit dialect differences.

3.3.7. Conclusions

The present experiment elaborated on the results of Experiment 1 by investigating the influence of animacy on auxiliary selection preferences. Such effects have been reported in the literature on unaccusativity in Italian. However, the present experiment failed to find animacy effects for German: both continuation of state and existence of state verbs show an identical auxiliary selection behavior for both animate and inanimate subjects.

The present experiment also provided a more fine-grained analysis of the change of state class. This class contains some verbs that exist in a prefixed and in a non-prefixed form (e.g., *modern*/*vermodern* “rot”). Our experimental results show that prefixing changes the auxiliary selection preference of these verbs; the prefixed form receives a telic interpretation and prefers *sein*, while the non-prefixed form allows both auxiliaries and seems to be ambiguous between a telic and an atelic reading.

Finally, we demonstrated that the uncontrolled process (involuntary reaction) class contains two subclasses, viz., verbs that imply motion (such as *torkeln* “tatter”) and ones that do not (such as *zittern* “jitter”). Verbs of the latter kind show a clear *haben* preference, while verbs of the former kind behave like motion verbs in that they allow both auxiliaries and exhibit dialect variation in their auxiliary selection preferences. This might indicate that verbs like *torkeln* “tatter” should be classified as members of the controlled process (motional) class, instead of as uncontrolled process verbs.

To summarize, the present experiment provided further evidence for the core/periphery distinction by demonstrating that peripheral verbs are subject to telicity effects. Animacy effects, another potential diagnostic for the core/periphery distinction, could not be demonstrated. Furthermore, we confirmed dialect variation a diagnostic for the core/periphery dichotomy: we

replicated the dialect effects found for the existence of state class in Experiment 1, and found an additional dialect effect for the change of state class.

3.4. Experiment 3: Effect of Telicity on Unaccusativity and Unergativity

Experiments 1 and 2 investigated unaccusative/unergative verbs with respect to auxiliary selection and impersonal passive formation. They supported the distinction of core vs. peripheral verbs based on evidence from gradient acceptability and dialectal variation. Experiment 2 also provided some initial evidence for telicity effects as a diagnostic of the peripheral status of a class. However, the results of Experiment 2 were limited to lexical telicity effects triggered by prefixing for certain verbs classes. The present experiment extends the investigation to telicity effects induced by syntactic factors, viz., by telic/atelic adverbials.

3.4.1. Introduction

This experiment investigates telicity effects for motion and emission verbs. The stimuli include directional and positional adverbials to induce a telic or atelic reading, as shown in (3.15) and (3.18) (repeated below).

(3.15) a. Der Zug ist/*hat in den Bahnhof gerumpelt.
the train is/has in the station rattled
"The train rattled into the station."

b. Der Zug *ist/hat im Bahnhof gerumpelt.
the train is/has into the station rattled
"The train rattled in the station."

(3.18) a. Die Frau ist/*hat ans Ufer geschwommen.
the woman is/has to the shore swam
"The woman swam to the shore."

b. Die Frau *ist/hat im Fluss geschwommen.
the woman is/has in the river swam
"The woman swam in the river."

To obtain plausible results, we have to make sure that an effect we might find is really due to the interaction of verb class and telicity. The mere presence of an adverbial might prompt subjects to vary their judgments, and thus cause the effect. For this reason, a control condition was included using peripheral unaccusative verbs. For the continuation of state and existence of state classes, stimuli involving two types of adverbials were constructed analogous to the ones used for motion and emission verbs. These adverbials varied in their aspectual properties: positional adverbials like *auf dem Rastplatz* "on the resting place" or *auf dem Beichtstuhl* "in the confessional" (see (3.7a) and (3.11a), repeated below) were contrasted with durational adverbials like

eine lange Zeit “for a long time” or *stundenlang* “for hours” (see (3.7b) and (3.11b), repeated below).

(3.7) a. Der Wanderer ?ist/hat auf dem Rastplatz verweilt.
the hiker is/has at the resting place stayed
“The hiker stayed at the resting place.”

b. Der Wanderer ?ist/hat eine lange Zeit verweilt.
the hiker is/has a long time stayed
“The hiker stayed a long time.”

(3.11) a. Die Betende ?ist/hat auf dem Beichtstuhl gekniet.
The praying person is/has on the confessional kneeled
“The praying person kneeled on the confessional.”

b. Die Betende ?ist/hat stundenlang gekniet.
The praying person is/has for hours kneeled
“The praying person kneeled for hours.”

If we fail to find a difference in auxiliary selection preference for the (a) and (b) stimuli in the control condition (see (3.7) and (3.11)), then this will be an indication that the telicity effect is genuine. On the other hand, the control condition can be used to confirm the dialectal variation obtained in Experiments 1 and 2, where a difference between speakers of northern and southern dialects was found for existence of state verbs, but not for continuation of state verbs.

3.4.2. Predictions

3.4.2.1. Constraints

The present experiment investigates how the auxiliary selection preference of a verb is affected by the telicity of the sentence, as induced by a telic or atelic adverbial. For the controlled process (motional) verbs and for uncontrolled process (emission) verbs (see (3.15) and (3.18)), we predict that a telic reading induces an auxiliary preference for *sein*, while an atelic reading induces a preference for *haben.* For the control condition (continuation of state and existence of state verbs), we predict that the choice of adverbial does not influence auxiliary preference (see (3.7) and (3.11)).

3.4.2.2. Constraint Types

In Experiments 1 and 2 we found dialectal differences for certain peripheral verb classes. We expect these differences to be replicated in the present experiment (for the controlled (motional) and existence of state classes).

For the emission and motion verbs, an interaction of telicity and auxiliary selection is predicted. An additional question is how the telicity effect interacts with the dialect preferences found for motion verbs in Experiments 1 and 2 (where explicit information about telicity was

absent). An intuitively correct prediction is that the telicity effect is strong enough to override dialect preferences in auxiliary selection. (Recall that this is what we found in Experiment 2 for change of state verbs, where prefixing induces a telic reading.)

3.4.3. Method

3.4.3.1. Subjects

Twenty-eight native Speakers of German from the same population as in Experiment 1 participated in the experiment. None of the subjects had previously participated in Experiment 1 or 2.

The data of two subjects were excluded because they were bilingual (by self-assessment). The data of another subject were excluded because he was a linguist (by self-assessment). The data of a forth subject were eliminated after an inspection of the responses showed that he had not completed the task adequately.

This left 24 subjects for analysis. Of these, 17 subjects were male, seven female; 23 subjects were right-handed, one left-handed. The age of the subjects ranged from 20 to 43 years, the mean was 26.9 years.

3.4.3.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials The experiment used two subdesigns. The first subdesign crossed the factors verb class (*Verb*), telicity (*Tel*), and auxiliary (*Aux*). The factor *Verb* had two levels, controlled process (motional) and uncontrolled process (emission). The factor *Tel* also had two levels, telic and atelic. These were realized by means of a directional PP or positional PP, as illustrated in examples (3.15) and (3.18). The factor *Aux* had two levels, *sein* and *haben.* This yielded a total of $Verb \times Tel \times Aux = 2 \times 2 \times 2 = 8$ cells. Eight lexicalizations were used for each of the cells, which resulted in a total of 64 stimuli. The lexicalizations for each class were the same as in Experiment 1 (see Table 3.2).

The second subdesign administered the control condition. It crossed the factors verb class (*Verb*), adverbial (*Adv*), and auxiliary (*Aux*). The factor *Verb* had two levels, continuation of state and existence of state. The factor *Adv* also had two levels, positional adverbials or durational adverbial, as illustrated in examples (3.7) and (3.11). This yielded a total of $Verb \times Tel \times Aux = 2 \times 2 \times 2 = 8$ cells. Eight lexicalizations were used for each of the cells, which resulted in a total of 64 stimuli. The lexicalizations for each class were the same as in Experiment 1.

A set of 24 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

The verb classes were matched for frequency using the same procedure as in Experiment 1.

3.4.3.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions We used the same instructions as in Experiment 1. Where contextualized stimuli were presented, subjects were told that each sentence would be presented in context, defined as a single sentence preceding the target sentence. Subjects were instructed to judge the acceptability of the target sentence, and to take the context into account in their judgments. The task was illustrated by examples.

Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

Eight test sets were used: each test set contained one lexicalization for each of the eight cells in the first subdesign, and one lexicalization for each of the eight cells in the second subdesign, i.e., a total of 16 items. Lexicalizations were assigned to test sets using a Latin square covering the full set of items.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 40 test items: 16 experimental items and 24 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to one of the test sets.

For each item, subjects were presented with the stimulus sentence and one context sentence that preceded it. This context sentence was meant to set the scene for the target sentence. Note that the present experiment did not manipulate context. However, Experiment 10, which manipulated context, was run as fillers for the present experiment. This made the change in experimental procedure necessary.

In the present experiment, all stimuli were presented in the same, neutral context: each sentence was preceded by the all focus question *Was gibt's neues?* "What's new?".

3.4.4. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

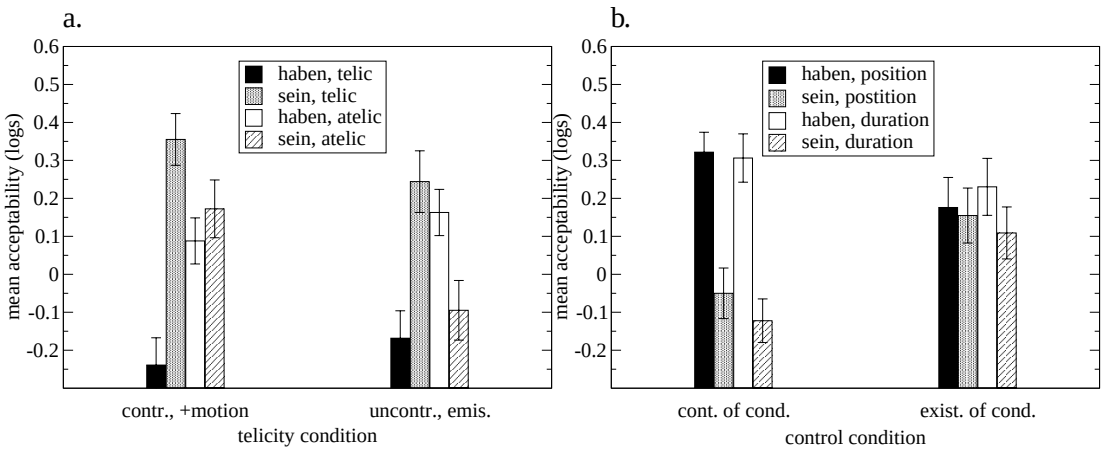

Figure 3.5: Interaction of telicity and auxiliary selection and control condition (Experiment 3)

3.4.4.1. Constraints

Telicity Condition The mean judgments for the telicity subexperiment are graphed in Figure 3.5a. An ANOVA revealed a main effect of *Aux* (auxiliary) that was significant by subjects and marginal by items ($F_1(1,23) = 18.812$, $p < .0005$; $F_2(1,7) = 3.666$, $p = .097$). The main effects of *Verb* (verb class) and *Tel* (telicity) failed to reach significance.

There was an interaction of *Aux* and *Verb*, which was significant by subjects only ($F_1(1,23) = 10.422$, $p = .004$; $F_2(1,7) = 2.636$, $p = .148$). Crucially, we found a highly significant interaction of *Aux* and *Tel* ($F_1(1,23) = 68.227$, $p < .0005$; $F_2(1,7) = 44.315$, $p < .0005$), confirming our prediction that telicity has an influence on auxiliary choice. All other interactions were non-significant.

Control Condition The mean judgments for the control condition are graphed in Figure 3.5b. An ANOVA revealed a significant main effect of *Aux* ($F_1(1,23) = 19.563$, $p < .0005$; $F_2(1,7) = 14.066$, $p = .007$). There were no main effects of *Verb* and *Adv* (adverbial).

However, there was an interaction of *Aux* and *Verb*, which was significant by subjects only ($F_1(1,23) = 24.716$, $p < .0005$; $F_2(1,7) = 2.841$, $p = .136$). Crucially, all interactions involving *Adv* were non-significant. This confirms our prediction that the type of adverbial has no influence on auxiliary choice in the control condition.

3.4.4.2. Constraint Types

As in Experiments 1 and 2, a re-analysis was performed with dialect as a between-subject factor. The criteria for assigning subjects to dialect areas were the same as in Experiments 1 and 2. There were 11 speakers of northern dialects, and 13 speakers of southern dialects.

The results of the by-dialect analysis are graphed in Figure 3.6 (note that auxiliary preferences are shown, not absolute judgments). For the telicity condition, we found a main effect

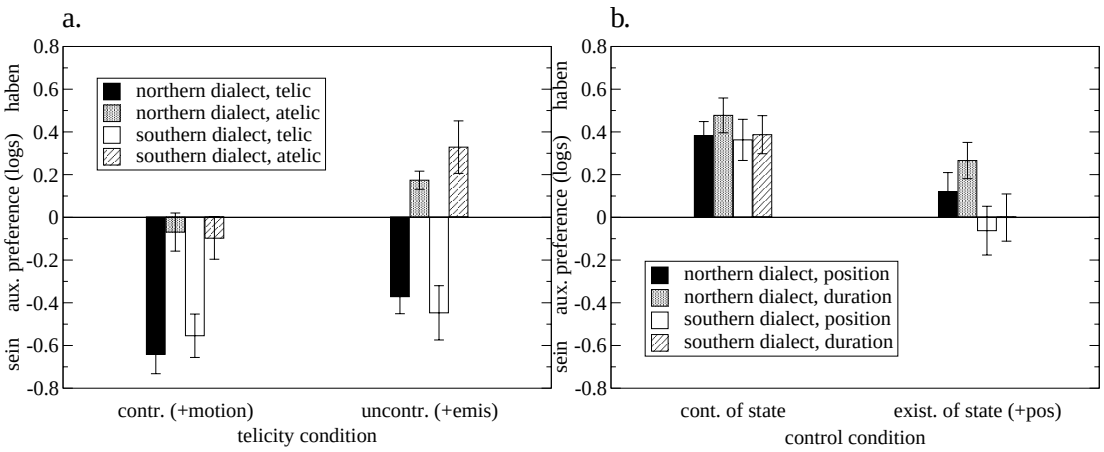

Figure 3.6: Interaction of telicity and auxiliary selection by dialect (Experiment 3)

of *Aux* ($F_1(1,22) = 18.095$, $p < .0005$) and interactions of *Aux* and *Verb* ($F_1(1,22) = 9.970$, $p = .005$) and of *Aux* and *Tel* ($F_1(1,22) = 65.716$, $p < .0005$). There was also an interaction of *Verb* and dialect ($F_1(1,22) = 9.711$, $p = .005$). This reflects the fact that speakers of northern and southern dialects differ in the strength of their auxiliary selection preferences for uncontrolled process (emission) verbs, but not for controlled process (motional) verbs (see Figure 3.6a). All other main effects and interactions were non-significant.

In the control condition, the by-dialect analysis revealed a main effect of *Aux* ($F_1(1,22) = 20.214$, $p < .0005$) and an interaction of *Aux* and *Verb* ($F_1(1,22) = 25.444$, $p < .0005$). All other main effects and interactions (including the ones involving dialect) failed to be significant.

3.4.5. Discussion

3.4.5.1. Constraints

As predicted, we found that auxiliary selection is sensitive to telicity for peripheral unergatives in the classes controlled process (motional) and uncontrolled process (emission). In both classes, a telic reading induces an auxiliary preference for *sein*, while an atelic reading induces a preference for *haben*, or at least a reduced preference of *sein* (see Figure 3.5a).

We failed to find an influence of type of adverbial on auxiliary selection in the control condition, which involved the peripheral unaccusative classes continuation of state and existence of state (see Figure 3.5b). This confirms that subjects are really reacting to the change in telicity induced by the adverbial, rather than making spurious distinctions between different types of adverbials.

3.4.5.2. Constraint Types

In Experiments 1 and 2, we discovered dialectal variation in the auxiliary selection behavior of peripheral, but not of core verbs. In the present experiment, we investigated peripheral verbs only, and expected dialectal differences consistent with those observed in the previous experiment. In the control condition, no significant dialect effects were found. Note that we had predicted a dialect effect for the existence of state class, based on the results of Experiments 1 and 2), where speakers of northern dialects preferred *haben*, while speakers of southern dialects judged *haben* and *sein* as equally acceptable. In the present experiment, only a non-significant tendency was observed (see Figure 3.6b).

An interesting observation concerning the effect of telicity on dialect preference can be arrived at by comparing the results of Experiment 1 with the results of the present experiment (see Figures 3.2 and 3.6a). For uncontrolled process (emission) verbs, subjects seem to assume an atelic reading in the absence of disambiguating information. This is true for speakers of both dialects (though the telicity effect is larger for speakers of southern dialects, which explains the interaction of verb class and dialect in the present experiment). However, for controlled process (motional) verbs, we observe an interesting dialect difference regarding auxiliary preferences. Speakers of northern dialects seem to assume a telic reading in the absence of disambiguating information (resulting in a *sein* preference), while speakers of southern dialects assume an atelic reading (resulting in a *haben* and *sein* being equally acceptable). However, explicit telicity information overrides these preferences in both dialects: in the atelic version, there is a clear *haben* preference, while *haben* and *sein* are equally acceptable in the atelic version (see again Figures 3.2 and 3.6a).

Note that this effect is analogous to the prefix effect we found in Experiment 2 for change of state verbs. Change of state verbs without a prefix are compatible with both a telic and an atelic reading. In northern dialects, the telic interpretation (selecting for *sein*) is preferred, while in southern dialect, the atelic interpretation (selecting for *haben*) is more acceptable. Once a prefix is added, however, only the telic interpretation is possible, and the dialect effect disappears (see Figure 3.4).

3.4.6. Conclusions

The present experiment investigated a subset of the verb classes from Experiments 1 and 2. The results replicated the dialect differences found in the earlier experiment, thus confirming that peripheral verbs are subject to dialect variation. Furthermore, we found that the auxiliary selection behavior of certain peripheral verbs is subject to telicity effects induced by sentential adverbials.

Taken together, Experiments 1–3 provide three criteria for the distinction between core and peripheral verbs:

- **Gradience** Core verbs show clear preferences for one auxiliary, while peripheral ones exhibit gradience, i.e., they allow both auxiliaries to a certain degrees.
- **Crosslinguistic Variation** The auxiliary selection preferences of core verbs are constant across languages (and dialects), while the preferences of peripheral verbs are subject to crosslinguistic or crossdialectal variation.
- **Telicity Effects** The auxiliary selection preferences of peripheral verbs are subject to telicity effects; for core verbs, no such effects are expected. Telicity can be induced by prefixing or by adverbials.

Note that the core/periphery distinction is based on a classification of verbs, not of constraints, which was what we were aiming for initially. However, there is an immediate connection between verb classes and types of constraints. Under the assumption that class membership is governed by a set of constraints, we can postulate that some of these constraints (call them hard constraints) determine the membership in core verb classes, while others (call them soft constraints) determine the membership in peripheral verb classes. (We will discuss this link between verb classes and constraint types in more detail in Section 4.1.2.)

In this setting, the distinction of core vs. peripheral verbs is a special case of the more general distinction of hard vs. soft linguistic constraints. Therefore, soft constraints are expected to cause gradient acceptability effects and are subject to telicity effects and crosslinguistic (or crossdialectal) variation. Hard constraints, on the other hand, are expected to induce binary acceptability judgments, and should be immune to telicity effects and stable across languages and dialects.

In Experiments 4–6, we will test the hard/soft distinction for three new phenomena: extraction, binding, and word order. We will also raise new questions regarding the interaction of hard and soft constraints. In the following chapter, Experiments 7–12 will investigate context effects on soft and hard constraints. Furthermore, we will return to the issue of crosslinguistic variation in Experiments 6 and 10–12.

3.5. Experiment 4: Extraction from Picture NPs

The results of Experiments 1–3 led to the hypothesis that linguistic constraints come in two types: soft and hard. Soft constraints (like the ones governing peripheral verb classes) induce gradient acceptability and are subject to crosslinguistic (crossdialectal) variation. Hard constraint (like the ones governing core verb classes), on the other hand, lead to binary acceptability and are immune to crosslinguistic (crossdialectal) differences.

The purpose of the present experiment is threefold. Firstly, it aims to validate the soft/hard dichotomy for a different syntactic phenomenon (extraction from picture NPs). Secondly, it provides data on constraint interaction (see Section 3.1.4) by investigating multiple

constraint violations. In particular, we will try to determine if soft and hard constraints differ with respect to multiple violations. Thirdly, the present experiment will provide data on the relative degree of unacceptability induced by the violation of six different constraints, based on which conclusions on constraint ranking can be drawn (see Section 3.1.2).

3.5.1. Background

The phenomenon under investigation is extraction from picture NPs, a construction for which gradient acceptability has been observed both in the theoretical (Erteschik-Shir 1981; Fiengo 1987; Kas 1991; Kluender 1992) and in the experimental literature (Cowart 1989a, 1997; Keller 1996a,b). The results on extraction obtained in this experiment will feed into the follow-up Experiment 5, which deals with the related phenomenon of binding in picture NPs.

Complex NPs are standardly assumed to be islands for extraction. Picture NPs, however, constitute well-known counterexamples to this assumption, as they allow island violations in certain cases. A number of factors are known to influence the island status of picture NPs. For instance, Kluender (1992) and Fiengo (1987) observe that definiteness has an influence on extractability: extraction from indefinite picture NPs is more acceptable than extraction from definite ones (see (3.27)).

(3.27) a. Which friend has Thomas painted a picture of?
b. ?Which friend has Thomas painted the picture of?

Extractability also depends on the aspectual class of the matrix verb. Vendler (1967) proposes to distinguish four aspectual classes: states, activities (unbounded processes), accomplishments (bounded processes), and achievements (point events). This classification can be further refined by taking into account the existential presupposition that some verbs carry (Diesing 1992). A verb like *tear up* presupposes the existence of its object, while a verb like *paint* carries no such presupposition. We will mark this presupposition using the feature [±EXISTENCE].

It has been observed (Diesing 1992; Erteschik-Shir 1981; Kluender 1992) that extraction from picture NPs is more acceptable for state verbs than for activity verbs (see (3.28)). For accomplishment and achievement verbs, a [−EXISTENCE] verb is more acceptable than a [+EXISTENCE] verb (see (3.29) and (3.30)).

(3.28) a. Which friend has Thomas owned a picture of?
b. ?Which friend has Thomas analyzed a picture of?
(3.29) a. Which friend has Thomas painted a picture of?
b. ?Which enemy has Thomas torn up a picture of?
(3.30) a. Which friend has Thomas found a picture of?
b. ?Which friend has Thomas lost a picture of?

A third factor influencing the acceptability of extraction from picture NPs is the referentiality of the extracted NP. It has been claimed (Kluender 1992) that referential NPs like *which friend* are more extractable than non-referential ones like *how many friends*:

(3.31) a. Which friend has Thomas painted a picture of?
b. ?How many friends has Thomas painted a picture of?

Previous experimental research has confirmed that all three factors influence the acceptability of extraction from picture NPs. Cowart (1997) demonstrated that indefinite picture NPs are easier to extract from than definite ones. Keller (1996a,b) replicated the definiteness effect, and also confirmed that verb class and referentiality influence the acceptability of extraction from picture NPs.

Previous research has investigated the effect of definiteness, verb class, and referentiality in isolation. The present experiment, in contrast, focuses on how these factors interact in picture NP extraction. Data on constraint interaction can be obtained by investigating the effect of multiple constraint violations on the degree of acceptability of a given structure (see our operational definition of constraint interaction in Section 3.1.4). To implement this approach, we adopt a constraint-based view of extraction from picture NP. We postulate the following set of constraints:

(3.32) **Constraints on Picture NPs**
a. DEFINITENESS (DEF): a picture NP has to be marked [−DEFINITE].
b. VERBCLASS (VERB): a verb subcategorizing for a picture NP has to be marked [−EXISTENCE].
c. REFERENTIALITY (REF): an NP extracted from a picture NP has to be marked [+REFERENTIAL].

Note that these constraints are purely descriptive. They reflect observations in literature on what constitutes a good picture NP (i.e., one from which extraction is allowed).

The second part of the present experiment deals with multiple violations of hard constraints. We investigate two constraints on *wh*-questions that intuitively seem to be hard constraints, in the sense of causing strong unacceptability when violated. The first constraint is INVERSION (INV) and states that in *wh*-questions, the subject and the auxiliary have to be inverted, as illustrated in example (3.33).

(3.33) a. Which friend has Sarah painted a picture of?
b. *Which friend Sarah has painted a picture of?

The second constraint on *wh*-extraction is called RESUMPTIVE (RES) and disallows resumptive pronouns, such as in the following example:

(3.34) a. Which friend has Sarah painted a picture of?
b. *Which friend Sarah has painted a picture of her?

Finally, we include a violation of number agreement as a control condition, consider example (3.35). This constraint, AGREEMENT (AGR), is not specific to extraction in the same way as the constraints on inversion and resumptive pronouns. Therefore, it can serve as a benchmark against which to compare the violation of these two constraints.

(3.35) a. Which friend has Sarah painted a picture of?
b. *Which friend have Sarah painted a picture of?

In terms of its acceptability pattern, this control is expected to cluster with the hard constraints on extraction (inversion and resumptive pronouns).

3.5.2. Introduction

This experiment has two subdesigns. The first one investigates soft constraints on picture NPs in (3.32), viz., DEFINITENESS, VERBCLASS, and REFERENTIALITY. We use stimuli like the ones in (3.27) to test violations of DEFINITENESS, while stimuli like the ones in (3.29) and (3.31) are used to test violations of VERBCLASS and REFERENTIALITY, respectively. Each stimulus can incur multiple violations; we include stimuli with a single violation of one of the three constraints, stimuli with two constraint violations, and stimuli that incur violations of all three constraints. This allows us to investigate constraint interaction, i.e., to determine whether constraint violations behave in an cumulative fashion.

The second subdesign deals with hard violations, i.e., with inversion, resumptive pronouns, and agreement. The stimuli were designed based on examples (3.33)–(3.35) in the previous section. Again, each stimulus can incur up to three constraint violations, which allows us to investigate the cumulativity of hard constraint violations.

3.5.3. Predictions

3.5.3.1. Constraints

In line with the claims in the theoretical literature, and with the results of previous experimental studies (Cowart 1989a, 1997; Keller 1996a,b), we predict a significant main effect of constraint violation for each of the soft constraints on extraction, i.e., for DEFINITENESS, VERBCLASS, and REFERENTIALITY (see (3.32)). Furthermore, we predict a significant main effect of constraint violation for the two hard constraint on extraction, i.e., INVERSION and RESUMPTIVE. We also expect a main effect of AGREEMENT violations (which was included as a control condition).

3.5.3.2. Constraint Ranking

In Section 3.1.2 we proposed an operational definition of constraint ranking based on the degree of unacceptability caused by a given constraint violation; the higher the degree of un-

acceptability caused by a violation, the more highly ranked the constraint. For the present experiment, this means that hard constraints are expected to be ranked higher than soft constraints: violations of DEFINITENESS, VERBCLASS, and REFERENTIALITY should produce a lesser degree of unacceptability than violations of the constraints INVERSION, RESUMPTIVE, and AGREEMENT. Such a pattern would be in line with the results of Experiments 1–3, where core verbs (governed by hard constraints) induced strong auxiliary selection preferences, while peripheral verbs (governed by soft constraints) were associated with weak tendencies.

A further question is how individual hard and soft constraints are ranked relative to each other. Are some soft constraint violations more serious than others? Intuitively, we would expect the answer to be yes, based on the diverse unacceptability pattern found for peripheral verbs in Experiment 1. The same question can be asked for hard constraints (but note that the core verb classes in Experiments 1–3 showed a uniformly binary auxiliary selection pattern). A set of planned comparisons will be used to compare the degree of unacceptability caused by individual soft and hard constraint violations.

3.5.3.3. Constraint Interaction

Another aspect of the present experiment is constraint interaction; we attempt to determine how multiple constraint violations affect the acceptability of a linguistic structure. Based on our operational definition of constraint interaction (see Section 3.1.4), diverse assumptions can be made about constraint interaction, leading to distinct predictions about the behavior of structures that incur multiple constraint violations.

Under an optimality theoretic approach the assumption is that constraint interaction is governed by the principle of strict domination, which states that the highest ranking constraint on which two structures conflict is crucial for deciding which of the structures is optimal. In the present experimental setting this means that a structure that incurs a violation of a constraint *C* should be less acceptable than any structure that only violates constraints that are ranked lower than *C*, even if it incurs multiple violations of such constraints.

An alternative approach to constraint interaction is to assume that violations are cumulative, i.e., the unacceptability of a structure increases directly with the number of constraints it violates. This means that the degree of unacceptability of a structure is simply the sum of all constraint violations it incurs.

The second question addressed by the present experiment is if soft and hard constraints differ with respect to multiple violations. It is conceivable that hard violations are subject to strict domination, while soft constraint violations are cumulative, or vice versa. The experiment comprises two subdesigns that deal with multiple violations of hard and soft constraint separately, and thus allows us to answer this question. A set of planned comparisons will be carried out to compare the degree of unacceptability caused by single, double, and triple violations of both hard and soft constraints.

3.5.4. Method

3.5.4.1. Subjects

Twenty-nine native speakers of English participated in the experiment. The subjects were recruited over the Internet by postings to relevant newsgroups and mailing lists. Participation was voluntary and unpaid. Subjects had to be linguistically naive, i.e., neither linguists nor students of linguistics were allowed to participate.

The data of two subjects were excluded because they were bilingual (by self-assessment). The data of a third subject were eliminated after an inspection of the responses showed that she had not completed the task adequately.

This left 26 subjects for analysis. Of these, 15 subjects were male, 11 female; three subjects were left-handed, 23 right-handed. The age of the subjects ranged from 17 to 52 years, the mean was 30.1 years.

3.5.4.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials The experiment included two subdesigns, one for soft constraints on extraction and one for hard constraints on extraction. The first subdesign dealt with soft constraint and crossed the factors *Def*, *Ref*, and *Verb*. The factor *Def* tested the constraint DEFINITENESS and had two levels (definite, indefinite, see (3.27)). The *Verb* tested the constraint VERBCLASS and also had two levels (accomplishment [−EXISTENCE], accomplishment [+EXISTENCE], see (3.29)). Similarly, the factor *Ref* had two levels (referential, non-referential, see (3.31)) and tested the constraint REFERENTIALITY. This yielded a total of $Def \times Ref \times Verb = 2 \times 2 \times 2 = 8$ cells.

The second subdesign dealt with hard constraints and crossed the factors *Inv*, *Res*, and *Agr*. There were two levels for *Inv* (inverted, non-inverted, see (3.33)), which tested the constraint INVERSION. The factor *Res* tested the constraint RESUMPTIVE and also included two levels (resumptive, no resumptive, see (3.34)). Finally, the factor *Agr* tested the constraint AGREEMENT, and also included two levels (number agreement, no number agreement, see (3.35)), yielding a total of $Inv \times Res \times Agr = 2 \times 2 \times 2 = 8$ cells. Four lexicalizations were used for each of the cells, which resulted in a total of 64 stimuli.

A set of 16 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

To control for possible effects from lexical frequency in the factor *Verb*, the two sets of lexicalizations of *Verb* ([+EXISTENCE] and [−EXISTENCE]) were matched for frequency. Frequency counts for the verbs and the head nouns were obtained from a lemmatized version of

the British National Corpus (90 million words of text, 10 million words of speech) and the average frequencies were computed for the lexicalizations of *wh*-phrase, subject NP, picture NP, and verb. An ANOVA confirmed that these average frequencies were not significantly different from each other.

3.5.4.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions We used an English version of the instructions in Experiment 1.

Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

Four test sets were used: each test set contained one lexicalization for each of the eight cells in the first subdesign, and one lexicalization for each of the eight cells in the second subdesign, i.e., a total of 16 items. Lexicalizations were assigned to test sets using Latin squares. A separate Latin square was applied for each subdesign.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 32 test items: 16 experimental items and 16 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to one of the test sets.

3.5.5. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

3.5.5.1. Constraints

Soft Constraints The mean judgments for soft constraint violations are graphed in Figure 3.7.[11] An ANOVA showed that the factor *Def* was significant by subjects, and marginal by items ($F_1(1,25) = 8.152$, $p = .009$; $F_2(1,3) = 7.199$, $p = .075$): extraction from indefinite picture NPs (mean $= .0448$) was more acceptable than extraction from definite ones (mean $= -.0051$). A main effect of *Ref* was also found ($F_1(1,25) = 14.612$, $p = .001$;

[11] This figures graphs *multiple* violations of soft and hard constraints, i.e., it compares the average acceptability of all structures that violate a given constraint *C* with the average acceptability of all structure that do not violate *C*. Some of these structure will incur violations of constraints other than *C*, and hence be of reduced acceptability.

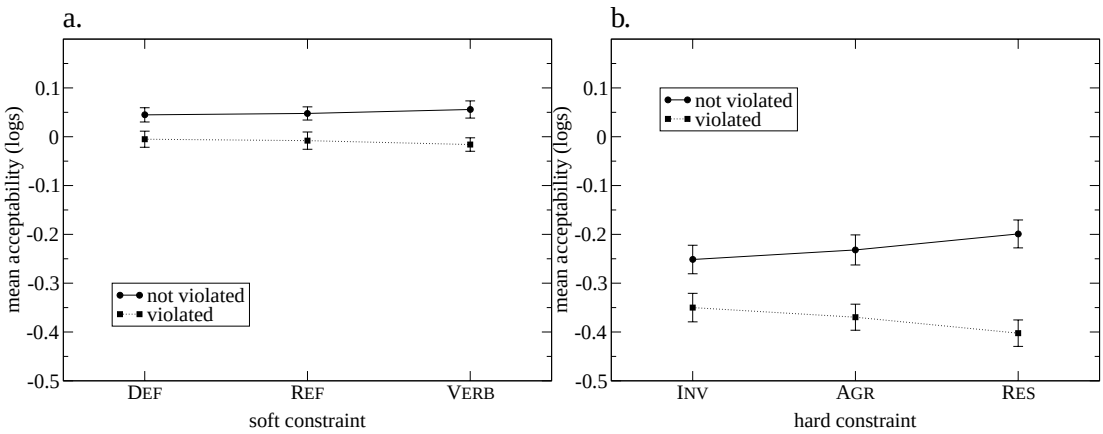

Figure 3.7: Comparison of soft and hard constraint violations, multiple violations (Experiment 4)

$F_2(1,3) = 11.765$, $p = .042$): extraction of referential *wh*-phrases (mean $= .0477$) was more acceptable than extraction of non-referential ones (mean $= -.0080$). Finally, there was a main effect of *Verb* ($F_1(1,25) = 17.075$, $p < .0005$; $F_2(1,3) = 17.234$, $p = .025$): verbs of the class $[-$EXISTENCE$]$ (mean $= .0558$) were more acceptable than $[+$EXISTENCE$]$ verbs (mean $= -.0160$). All interactions failed to be significant.

Hard Constraints The mean judgments for hard constraint violations are graphed in Figure 3.7.[12] An ANOVA revealed a significant main effect of *Inv* ($F_1(1,25) = 12.148$, $p = .002$; $F_2(1,3) = 14.475$, $p = .032$): inverted *wh*-questions (mean $= -.2515$) were significantly more acceptable than uninverted ones (mean $= -.3500$). A main effect of *Res* was also found ($F_1(1,25) = 37.115$, $p < .0005$; $F_2(1,3) = 17.568$, $p = .025$): *wh*-questions without resumptives (mean $= -.1991$) were more acceptable than ones with resumptives (mean $= -.3500$). Finally, a main effect of *Agr* was found ($F_1(1,25) = 23.472$, $p < .0005$; $F_2(1,3) = 26.948$, $p = .014$): stimuli with number agreement (mean $= -.2319$) were more acceptable than the ones without (mean $= -.3697$).

There was a significant interaction between *Inv* and *Res* ($F_1(1,25) = 9.962$, $p = .004$; $F_2(1,3) = 16.287$, $p = .027$), and an interaction of *Res* and *Agr*, which however was significant only by subjects ($F_1(1,25) = 9.285$, $p = .005$; $F_2(1,3) = 2.566$, $p = .207$). The interaction of *Inv* and *Agr* was non-significant, as was the three-way interaction of all factors.

3.5.5.2. Constraint Ranking

We carried out a series of planned comparisons to determine if there are differences in the ranking of constraints. We compared the degree of unacceptability caused by single constraint

[12] Again, this figures graphs multiple constraint violations.

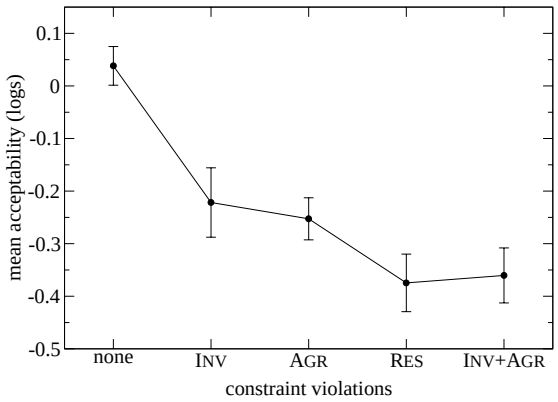

Figure 3.8: Constraint ranking, single violations (Experiment 4)

violations (graphed in Figure 3.8). Three planned comparisons were carried for the first subexperiment (soft constraints), hence the significance level was set at $p = .0167$ (Bonferroni adjustment). The second subexperiment also comprised three planned comparisons, hence we again set $p = .0167$.[13]

First we compared the degree of unacceptability caused by single violations of the soft constraints VERB (mean $= .0540$), REF (mean $= .0877$), and DEF (mean $= .0473$). None of the comparisons yielded a significant difference. Then we carried out planned comparisons on single violations of the hard constraints INV (mean $= -.2217$), AGR (mean $= -.2527$), and RES (mean $= -.3746$). We found that a RES violation was significantly more serious than an AGR violation (by subjects only, $F_1(1,25) = 9.540$, $p = .005$; $F_2(1,3) = 8.327$, $p = .063$). Also, a RES violation was marginally more serious than an INV violation (by subjects only, $F_1(1,25) = 5.744$, $p = .024$; $F_2(1,3) = 2.424$, $p = .217$). There was no significant difference between an AGR and an INV violation.

[13] What follows is a general remark on how planned comparisons are handled in this thesis. We use the Bonferroni method to reduce the risk of a Type I error: if c comparisons are carried out on same data, then a significance level of p/c is used (where $p = .05$). However, we carry out separate Bonferroni adjustments for sets of comparisons that are orthogonal (i.e., statistically independent; see Hays 1964: Ch. 14 of issues relating to orthogonality in planned comparisons). This strategy is less conservative than performing an overall Bonferroni adjustment for all comparisons in the experiment, which would increase the risk of a Type II error.

In the present experiment, for example, we set $c = 3$ when we compare the ranks of the three single soft violations (these are three non-orthogonal comparisons). In a second set of tests, we compare single, double, and triple violations, and again set $c = 3$ (another three non-orthogonal comparisons). These two sets of comparisons are orthogonal, which justifies the use of separate Bonferroni adjustments, instead of using an overall adjustment of $c = 6$. The same situation occurs for the two sets of comparisons carried out for the second subexperiment. Also note that the comparisons for the two subexperiments are orthogonal, hence separate Bonferroni adjustments can be used.

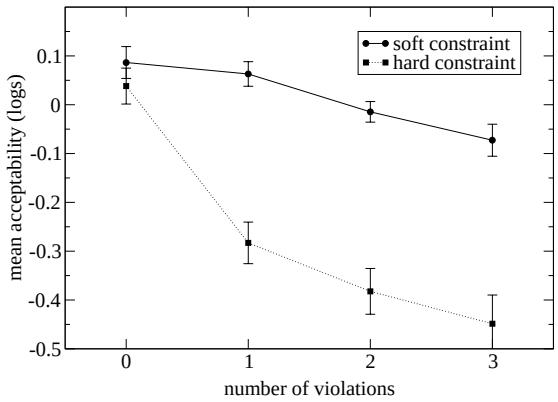

Figure 3.9: Cumulativity of constraint violations (Experiment 4)

3.5.5.3. Constraint Interaction

To test the hypothesis that constraint violations are cumulative, we carried out another series of planned comparisons. We determined if there is a significant difference between the acceptability of structures with zero, one, two, and three constraint violations. For this, we computed the mean acceptability of sentences with zero violations (one sentence type), one violation (mean of three sentence types, as there were three constraints per subexperiment), two violations (mean of three sentence types for all combinations of two constraint violation), and three violations (one sentence type). The resulting mean acceptability scores are graphed in Figure 3.9 for both soft and hard constraints. Three planned comparisons where carried out to test for the cumulativity for soft violations, hence the significance level was set at $p = .0167$ (Bonferroni adjustment). Another set of three comparisons was carried out for hard constraints, again resulting in a significance level of $p = .0167$.[14]

For the soft constraints, the difference between zero violations (mean $= .0865$) and one violation (mean $= .0630$) failed to be significant, while the difference between one violation and two violations (mean $= -.0146$) was significant by subjects and marginal by items ($F_1(1,25) = 10.685$, $p = .003$; $F_2(1,3) = 14.646$, $p = .031$). The difference between two violations and three violations (mean $= -.0628$) again failed to reach significance.

For the hard constraints, there was a significant difference between zero violations (mean $= .0382$) and one violation (mean $= -.2830$) ($F_1(1,25) = 27.869$, $p < .0005$; $F_2(1,3) = 60.338$, $p = .004$). The difference between a single violation and a double violation (mean $= -.3814$) was significant by subjects only ($F_1(1,25) = 16.552$, $p < .0005$; $F_2(1,3) = 10.893$, $p = .046$). We failed to find a significant difference between two violations and three violations (mean $= -.4486$).

We carried out another set of tests to determine if the principle of strict domination

[14] See Footnote 13 on how planned comparisons are handled in this thesis.

was instantiated in the experimental data. Recall that strict domination means that a structure that incurs a violation of a constraint C is less acceptable than any structure that only violates constraints that are ranked lower than C, even if this structure incurs multiple violations of lower ranked constraints. We have already established that RES outranks INV and AGR, i.e., that a violation of RES is more serious than a violation of either INV or AGR (see Figure 3.8). Under strict domination, we would now expect that a violation of RES (mean $= -.3746$) is more serious than even a combined violation INV and AGR (mean $= -.3604$). However, a post-hoc test comparing the acceptability of these two conditions failed to be significant (see also Figure 3.8). This post-hoc test used a significance level of $p = .0167$.[15]

While this result indicates that hard constraints do not interact according to the principle of strict domination, it is still possible that soft constraints are strictly dominated by hard ones. To test this, we conducted a post-hoc comparison of a single hard violation (mean $= -.2830$) with a triple soft violation (mean $= -.0727$); the difference in acceptability was significant ($F_1(1,25) = 20.096$, $p < .0005$; $F_2(1,3) = 24.115$, $p = .016$), and is illustrated in Figure 3.9. Again a significance level of $p = .0167$ was assumed for this post-hoc test.

3.5.6. Discussion

3.5.6.1. Constraints

We found that violations of soft constraints such as DEFINITENESS, VERBCLASS, and REFERENTIALITY lead to a significant decrease of the acceptability of extraction from picture NPs. This result provides an experimental confirmation of relevant claims in the theoretical literature, which typically rely on intuitive data.

We also investigated a set of hard constraints on *wh*-extraction: inversion, resumptive pronouns, and agreement. As expected, a violation of any of these constraints significantly decreases the acceptability of *wh*-extraction.

3.5.6.2. Constraint Ranking

In the light of the results from Experiments 1–3, we predicted that hard constraints are ranked higher than soft ones, i.e., hard violations cause a higher degree of unacceptability than soft violations. This prediction was borne out by the experimental results. It seems that violations of soft constraints cause only a mild decrease in acceptability, while violations of hard constraints lead to serious unacceptability (see Figure 3.7).

With respect to the ranking of individual constraints, we failed to find a difference between the degree of unacceptability incurred by the three soft violations. For hard constraints,

[15] This test constitutes a post-hoc test as it is based on the results of the planned comparisons that were carried out to determine constraint ranking. We used the same significance level for the post-hoc test and for the associated planned comparison, i.e., we set $p = .0167$.

however, an effect of constraint type was obtained: a violation of RES is more serious than violations of INV or AGR (see Figure 3.8). This result indicates that gradient acceptability is not limited to a particular constraint type; rather, gradience occurs with both hard constraints (as evidenced by the present experiment) and soft constraints (as evidenced by Experiments 1–3).

3.5.6.3. Constraint Interaction

We found evidence for the hypothesis that constraint violations are cumulative: the more constraints a structure violates, the higher its degree of unacceptability. This finding holds for both soft and hard violations (see Figure 3.9).

We also showed that even a single hard violation can induce a higher degree of unacceptability than three soft violations. This finding is compatible with the concept of strict domination if we assume that hard constraints dominate soft ones. However, it is also compatible with an cumulative scheme of constraint interaction under the assumption that the combined unacceptability associated with three soft violations is smaller than the unacceptability associated with a single hard violation (see Figure 3.9).

Furthermore, we found evidence against strict domination among hard constraints. The constraint RES is ranked higher than both INV and AGR. However, the combined violations of INV and AGR are as unacceptable as a single violation of RES (see Figure 3.8). Such a ganging up of constraint violations should be impossible under strict domination; the combination of two lower ranked violations should not compensate for a single violation of a higher ranked constraint. Under an cumulative constraint combination scheme, on the other hand, such ganging up effects are easily accounted for: the combined unacceptability associated with the violation of two lower ranked constraints is equal to the unacceptability associated with the violation of a single higher ranked constraint.

3.5.7. Conclusions

Based on the results of Experiments 1–3 we hypothesized that soft constraints cause gradient acceptability effects, while hard constraints induce binary acceptability judgments.

In the present experiment, however, we found evidence for gradient acceptability in hard constraint violations, disconfirming the initial hypothesis that gradience is limited to soft constraints. On the other hand, the data show that soft violations lead to a significantly lesser degree of unacceptability than hard ones. In general, soft violations seem to be associated with mild unacceptability, while hard constraint violations trigger strong unacceptability. The fact that hard violations are seriously unacceptable might lead to the intuitive perception of hard constraints as binary: it is difficult to detect gradience in seriously unacceptable structures unless one makes use of experimentally collected judgment data that allow fine distinctions in acceptability.

As for constraint interaction, the evidence suggests that constraint violations are cumulative, for both hard and soft constraints. Also, we found evidence for the ganging up of constraint violations, which is unexpected under an OT-type strict domination scheme. This serves as initial evidence against an OT-type model of constraint interaction, at least under the operational interpretation of strict domination that was put forward in Section 3.1.4.

3.6. Experiment 5: Exempt Anaphors and Picture NPs

Experiments 1–3 dealt with constraints types. They lead to a classification of constraints into soft and hard ones based on the observation that soft constraints cause mild unacceptability and are subject to crosslinguistic (or crossdialectal) variation. Hard constraints, on the other hand, fail to exhibit these effects and induce strong unacceptability when violated. Experiment 4 investigated constraint ranking and constraint interaction and showed that constraints are ranked, leading to the preliminary conclusion that constraint violations are cumulative. The present experiment aims to extend the study of constraint ranking and constraint interaction to a new, though related phenomenon: binding of anaphors and pronouns in picture NPs.

3.6.1. Background

3.6.1.1. Binding Theory

Binding theory is the module of grammar that regulates the interpretation of noun phrases (NPs). Three types of noun phrases are generally distinguished: (a) anaphors, i.e., reflexives such as *herself*, and reciprocals such as *each other*, (b) pronouns such as *he* and *her*, and (c) referring expressions such as *Hanna* or *the woman*,

The task of binding theory is to determine which noun phrase can be *coreferential*, i.e., refer to the same individual. Coreference is normally indicated using subscripts:

(3.36) a. Hanna$_i$ admires *her$_i$/herself.
b. Hanna$_i$ thinks that Peter admires her$_i$/*herself$_i$.
c. *She$_i$ admires Hanna$_i$.

In example (3.36a), the proper name *Hanna* and the pronoun *her* cannot refer to the same person, i.e., they cannot be coreferential (as indicated by the "*"). The pronoun cannot be *bound* by the proper name. In (3.36b), on the other hand, *Hanna* is a potential binder for *her*, i.e., coreference is possible. The situation for the reflexive is exactly opposite; *Hanna* and *herself* can be coreferential in (3.36a), but not in (3.36b).

There are structural conditions that determine the binding possibilities of anaphors and pronouns. Principle A of binding theory captures the binding requirements for anaphors; it states that an anaphor has to be bound within a certain local domain (Chomsky 1986). Principle B, on the other hand, states that pronouns cannot be bound within its local domain. It

follows that anaphors and pronouns are in complementary distribution, i.e., anaphors can be bound when pronouns cannot be bound, and vice versa. Principle C of binding theory deals with referring expressions (such as proper names); it requires that a referring expression must not be bound, and thus rules out sentences like (3.36c).

Binding theoretical issues have mainly been addressed by theoretical linguists. However, a small experimental literature exists, including a series of experiments by Gordon and Hendrick (1997, 1998a,b), who focused on native speakers' judgments of the coreference of proper names and pronouns. Their results provided evidence for Principle B and its formulation terms of Chomsky's (1986) notion of c-command. However, Gordon and Hendrick found only limited evidence for the validity of Principle C. (Experiment 14 reports a replication of some of Gordon and Hendrick's (1997) studies and provides a more detailed account of their findings.) Another relevant experimental study is reported by Cowart (1997), who investigated the binding properties of anaphors and demonstrated that an anaphor can be bound by a remote antecedent (contrary to the requirements of Principle A) if the anaphor occurs inside a coordinated NP. No previous experimental study has investigated the behavior of exempt anaphors, which present a problem to most formulations of Principles A and B and therefore have generated great theoretical interest. Exempt anaphora are the subject of the present experiment.

3.6.1.2. Exempt Anaphors

It has been observed by a number of authors (e.g., Pollard and Sag 1994; Reinhart and Reuland 1993) that in certain configurations, anaphors are exempt from binding theory. In such cases, the anaphor is not subject to Principle A. Relevant configurations include picture NPs without possessors, as illustrated in (3.37a), where the binding of an anaphor and a pronoun are both acceptable. When there is a possessor in the picture NP, the relevant domain for anaphoric binding is the NP, and anaphors are claimed to be unacceptable in sentences like (3.37b), while pronouns are fine.

(3.37) a. Hanna$_i$ found a picture of her$_i$/herself$_i$.
b. Hanna$_i$ found Peter's picture of her$_i$/*herself$_i$.

On the basis of such examples, authors like Pollard and Sag (1994) have argued that Principle A should be formulated so as not to apply to anaphors in sentences such as (3.37a). The assumption is that the binding properties of such anaphors are governed by non-syntactic factors, including processing and discourse constraints.

The present study has a double purpose. First, we attempt to clarify the empirical status of exempt anaphors. By conducting a study with linguistically naive native speakers we can determine whether anaphors and pronouns are perceived as equally acceptable in configurations like the one in (3.37a).

The second purpose is to shed light on the factors that influence the distribution of pronouns and exempt anaphors. In Experiment 4 we identified a set of factors that have an effect on extraction from picture NPs (referentiality, definiteness, aspectual class of the matrix verb). Our working hypothesis is that these factors also influence binding in picture NPs. If correct, this hypothesis would entail that binding and extraction should receive a unified linguistic account (the two phenomena have traditionally been treated separately).

3.6.2. Introduction

This experiment has two subdesigns. The first one investigates how the exempt status of an anaphor is influenced by the definiteness of the picture NP and by the aspectual class of the matrix verb. As an example of definiteness consider the minimal pair in (3.38): the picture NP in (3.38a) is indefinite and the one in (3.38b) is definite.

(3.38) a. Hanna$_i$ found a picture of her$_i$/herself$_i$.
b. Hanna$_i$ found the picture of her$_i$/herself$_i$.

The factor verb class is illustrated in example (3.39): *find* and *lose* are examples of achievement verbs, while *take* and *destroy* are accomplishment verbs; *find* and *take* are [−EXISTENCE], while *lose* and *destroy* are [+EXISTENCE] (see Section 3.5.1 for an explanation of these verb classes).

(3.39) a. Hanna$_i$ found a picture of her$_i$/herself$_i$.
b. Hanna$_i$ lost a picture of her$_i$/herself$_i$.
c. Hanna$_i$ took a picture of her$_i$/herself$_i$.
d. Hanna$_i$ destroyed a picture of her$_i$/herself$_i$.

The second subexperiment was designed to test the influence of an intervening NP, as illustrated by the minimal pair in (3.40). The intervention of a potential binder was identified by both Asudeh (1998) and Pollard and Sag (1994) as a relevant factor in determining the exempt status of an anaphor. According to Pollard and Sag (1994), the anaphor in (3.40a) is exempt because it does not have a potential referential binder in its local domain (the picture NP), whereas the anaphor in (3.40b) is not exempt since the picture NP contains a local referential nominal. The second subexperiment also tested the influence of the referentiality of the binder, as illustrated in (3.41). We also included a control condition where the intervening NP is the binder, as shown by the minimal pairs in (3.42):

(3.40) a. Hanna$_i$ found a picture of her$_i$/herself$_i$.
b. Hanna$_i$ found Peter's picture of her$_i$/herself$_i$.
(3.41) a. Hanna$_i$ found Peter's picture of her$_i$/herself$_i$.
b. The woman$_i$ found Peter's picture of her$_i$/herself$_i$.
c. Each woman$_i$ found Peter's picture of her$_i$/herself$_i$.

(3.42) a. Hanna$_i$ found Peter's picture of her$_i$/*herself$_i$.
b. Hanna found Peter's$_i$ picture of *him$_i$/himself$_i$.

In the present experiment, we elicited acceptability judgments for both the anaphor and the pronoun in configurations like the ones in (3.38)–(3.42). Our aim is to test if the factors definiteness, verb class, referentiality, and the intervention of a binder have a significant influence on the binding theoretic status of a given configuration.

3.6.3. Predictions

3.6.3.1. Constraints

In line with the binding literature, we predict that an anaphor and a pronoun are equally acceptable in examples like (3.37a). This means that we should fail to find a main effect of NP type (anaphor or pronoun). Furthermore, we expect that the intervention of a potential binder (see (3.40)) influences the exempt status of an anaphor, in line with the theoretical claims by Asudeh (1998) and Pollard and Sag (1994). Hence we should find a significant interaction of intervention and NP type.

While previous experimental studies showed that referentiality can affect binding (Gordon and Hendrick 1998b), there is no previous experimental work dealing specifically with exempt anaphors or with factors such as definiteness and verb class. However, there is some discussion of such effects in the theoretical literature (Chomsky 1986; Kuno 1987; Pollard and Sag 1994; Reinhart and Reuland 1993), which would lead us to predict that referentiality, definiteness, and verb class to influence binding in picture NPs. This means that our experiment should show interactions between NP type and these three factors.

Finally, Principle A predicts that anaphors lose their exempt status in the control condition (see (3.42)), where there is a referential potential local binder inside the picture NP. For the indicated coreference, binding theory predicts that (3.42a) should be unacceptable with the anaphor and acceptable with the pronoun, while (3.42b) is acceptable with the anaphor and unacceptable with the pronoun. This should manifest itself in the experiment as an interaction of binder and NP type.

3.6.3.2. Constraint Ranking

Under the assumption that binding and extraction in picture NPs are governed by similar constraints, we expect the constraints on definiteness, referentiality, and verb class to be soft constraints. This means that they should induce only small changes in the acceptability of anaphors or pronouns. On the other hand, the intervention of another potential binder should have a stronger influence on the exempt status of an anaphor. It should trigger an effect characteristic of a hard constraint violation. Planned comparisons will be used to compare the difference in acceptability caused by violations of soft and hard constraints.

3.6.3.3. Constraint Interaction

As in Experiment 4 we expect constraint violations to be cumulative. Also, we expect to find evidence for the ganging up of lower ranked constraints against a higher ranked one. Again, planned comparisons will be carried out to test these predictions.

3.6.4. Method

3.6.4.1. Subjects

Fifty-eight native Speakers of English from the same population as in Experiment 4 participated in the experiment. None of the subjects had previously participated in either Experiment 4.

The data of one subject were excluded because she was a linguist (by self-assessment). The data of five subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 52 subjects for analysis. Of these, 24 subjects were male, 28 female; four subjects were left-handed, 48 right-handed. The age of the subjects ranged from 17 to 57 years, the mean was 28.7 years.

3.6.4.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials The experimental materials included two subdesigns. The first subdesign used the factors *Def*, *Verb*, and *Ana*. The factor *Def* tested the effect of the constraint DEFINITENESS and had two levels (definite, indefinite, see (3.38)). The factor *Verb* tested the effect of VERBCLASS and had three levels (achievement [−EXISTENCE], accomplishment [−EXISTENCE], accomplishment [+EXISTENCE], see (3.39a), (3.39c), (3.39d)). The factor *Ana* tested the effect of NP type and had two levels (anaphor or pronoun). This yielded a total of $Def \times Verb \times Ana = 3 \times 2 \times 2 = 12$ cells.

The second subdesign included the factors *Ref*, *Bind*, and *Ana*. The factor *Ref* tested the constraint REFERENTIALITY and had three levels (proper name, definite NP, quantified NP, see (3.41)). The factor *Bind* tested the effect of type of binder and had two levels (remote or local binder, see (3.42)). To test the effect of NP type, the factor *Ana* included two levels (anaphor, pronoun). This yielded a total of $Ref \times Bind \times Ana = 3 \times 2 \times 2 = 12$ cells. In both subexperiments, four lexicalizations were used for each of the cells, which resulted in a total of 96 stimuli.

A set of 24 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

The lexicalizations were matched for frequency using the same procedure as in Experiment 4.

3.6.4.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions We used a modified English version of the instructions in Experiment 1. Subjects were instructed to judge the acceptability of coreference. This was defined as follows: "Your task is to judge how acceptable each sentence is by assigning a number to it. By acceptability we mean the following: Every sentence will contain two expressions in ALL CAPITALS. A sentence is acceptable if these two expressions can refer to the same person." The task was illustrated by examples.

Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

Four test sets were used: each test set contained one lexicalization for each of the 12 cells in the first subdesign, and one lexicalization for each of the 12 cells in the second subdesign, i.e., a total of 24 items. Lexicalizations were assigned to test sets using Latin squares. A separate Latin square was applied for each subdesign.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 48 test items: 24 experimental items and 24 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to one of the test sets.

3.6.5. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

3.6.5.1. Constraints

Verb Class and Definiteness The ANOVA on the first subexperiment yielded a main effect of *Verb* (verb class) ($F_1(2,102) = 9.345$, $p < .0005$; $F_2(2,6) = 4.839$, $p = .056$): [−EXISTENCE] accomplishment verbs like *take* were significantly less acceptable (mean = .3715) than [+EXISTENCE] accomplishment verbs like *destroy* (mean = .4653) or [−EXISTENCE] achievement verbs like *find* (mean = .4616). The main effect of *Def* (definiteness) was small and only significant by subjects ($F_1(1,51) = 7.927$, $p = .007$; $F_2(1,3) = 1.207$, $p = .352$). Definite

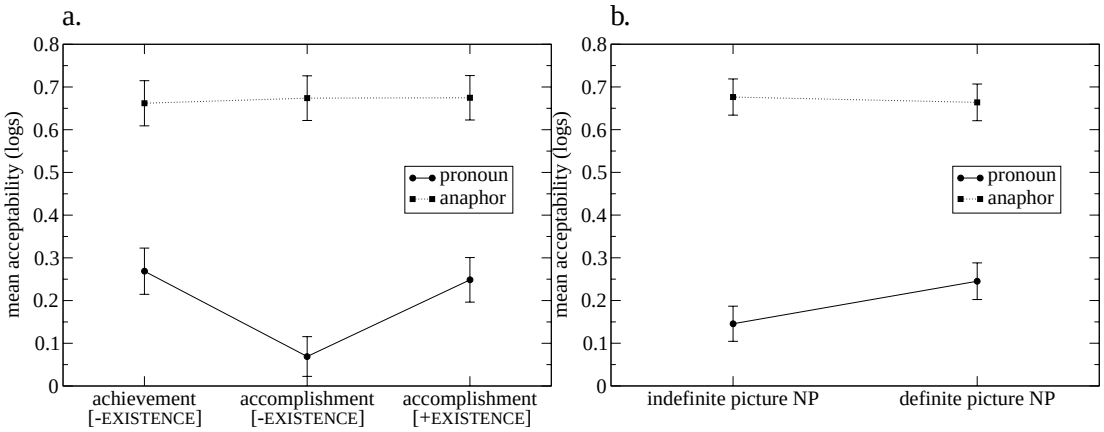

Figure 3.10: Interactions of *Verb* and *Ana*, and of *Def* and *Ana*, multiple violations (Experiment 5)

picture NPs (mean = .4546) were more acceptable than indefinite ones (mean = .4110). We also found a large and highly significant main effect of *Ana* (NP type) ($F_1(1,51) = 137.471$, $p < .0005$; $F_2(1,3) = 105.005$, $p = .002$). Anaphors (mean = .6702) were more acceptable than pronouns (mean = .1954).

The ANOVA also revealed a significant interaction of *Verb* and *Ana* ($F_1(2,102) = 11.275$, $p < .0005$; $F_2(2,6) = 6.193$, $p = .035$). This interaction is graphed in Figure 3.10a, which shows that there is a decrease in the acceptability of pronouns for $[-\text{EXISTENCE}]$ accomplishment verbs. An interaction of *Def* and *Ana* was also found, which however was significant by subjects only ($F_1(1,51) = 11.849$, $p = .001$; $F_2(1,3) = 2.168$, $p = .237$). Figure 3.10b shows that the acceptability for pronouns is increased for definite picture NPs. The interaction of *Verb* and *Def*, as well as the three-way interaction of *Verb*, *Def*, and *Ana* failed to be significant.

Binder and Referentiality The ANOVA on the second subexperiment revealed a main effect of *Bind* (remote or local binder), which however was significant by subjects only ($F_1(1,51) = 7.851$, $p = .005$; $F_2(1,3) = 4.284$, $p = .130$). A remote binder (mean = .4816) was more acceptable than a local binder (mean = .4085). The factor *Ref* (referentiality) was highly significant ($F_1(2,102) = 68.244$, $p = .001$; $F_2(2,6) = 12.197$, $p = .008$); quantified binders like *each woman* (mean = .4008) were less acceptable than non-quantified binders such as *Hanna* (mean = .4672) or *the woman* (mean = .4670). Finally, we replicated the effect of *Ana* found in the first subexperiment ($F_1(1,51) = 68.244$, $p < .0005$; $F_2(1,3) = 45.725$, $p = .007$). Again, anaphors (mean = .5800) were more acceptable than pronouns (mean = .3101).

The ANOVA also demonstrated a significant interaction of *Bind* and *Ref* ($F_1(2,102) = 3.966$, $p = .022$; $F_2(2,6) = 10.638$, $p = .011$). The interaction of *Bind* and *Ana* was significant by subjects and marginal by items ($F_1(1,51) = 35.051$, $p < .0005$; $F_2(1,3) = 6.274$, $p = .087$).

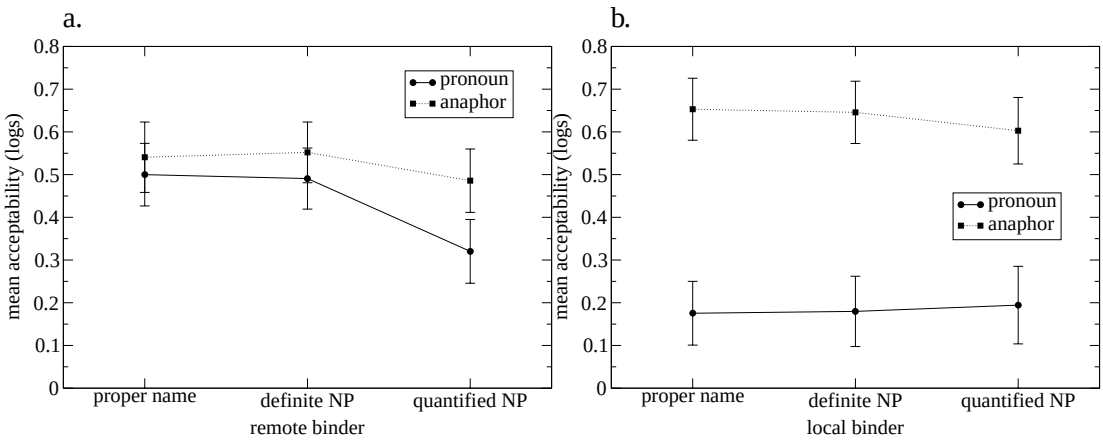

Figure 3.11: Interaction of *Ref*, *Ana* and *Bind*, multiple violations (Experiment 5)

Finally, a three-way interaction of *Bind*, *Ref*, and *Ana* was also obtained (significant by subjects and marginal by items, $F_1(2,102) = 4.041$, $p = .020$; $F_2(2,6) = 4.543$, $p = .063$). This interaction is graphed in Figure 3.11. An inspection of Figure 3.11a shows that in the remote binder condition, pronouns and anaphors are equally acceptable if the binder is a proper name or a definite NP. However, if the binder is a quantified NP, the acceptability for pronouns decreases. There is no such effect in the control condition (local binder, see Figure 3.11b). A post-hoc Tukey test on the *Bind*/*Ref*/*Ana* interaction confirms this observation: for the remote binder condition, the difference between pronoun and anaphor is not significant for proper names and definite NPs, but reaches significance for the quantified NPs (by subjects only, $\alpha < .01$). For the local binder condition, on the other hand, the difference between pronoun and anaphor is significant for all three binders ($\alpha < .05$).

A comparison of Figures 3.10 and 3.11 shows that picture NPs only have exempt status if there is an intervening potential binder, i.e., in the remote binder condition. If there is no intervening binder, pronouns are highly unacceptable with picture NPs—we get essentially the same acceptability pattern as in the case of a local binder. To confirm this observation, we conducted an ANOVA on the data that overlapped from the two subexperiments (see (3.40) for an example). The factors were *Ana* (pronoun or anaphor) and the new factor *Int*, which had two levels (intervening potential binder or not). The factor *Int* tested the additional constraint INTERVENE (INT), which penalizes the existence of an intervening potential binder. A main effect of *Int* was found, which however was significant by subjects only ($F_1(1,51) = 5.142$, $p = .028$; $F_2(1,3) = 1.747$, $p = .278$). We also found a main effect of *Ana*, which was significant by subjects and marginal by items ($F_1(1,51) = 33.181$, $p < .0005$; $F_2(1,3) = 6.987$, $p = .077$). Crucially, there was a significant interaction of *Int* and *Ana* ($F_1(1,51) = 35.432$, $p < .0005$; $F_2(1,3) = 15.608$, $p = .029$). This interaction is graphed in Figure 3.12.

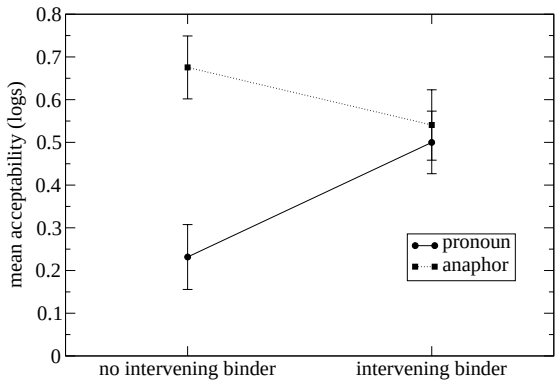

Figure 3.12: Interaction of *Int* and *Ana*, single violations (Experiment 5)

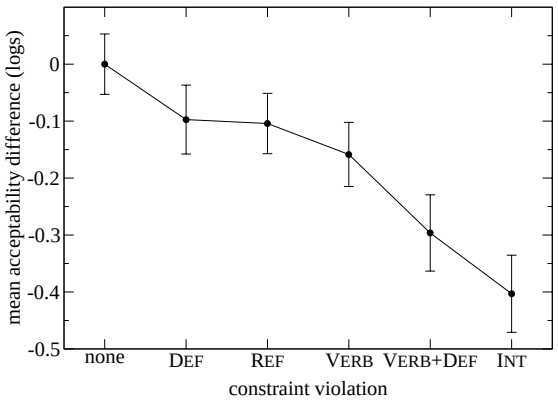

Figure 3.13: Constraint ranking, single violations (Experiment 5)

3.6.5.2. Constraint Ranking

As in Experiment 4, a series of planned comparisons was carried out to determine constraint rankings. We compared the change in binding preference caused by single violations of the soft constraints DEF, REF, and VERB with the change in binding preference caused by a single violation of the constraint INT. Figure 3.13 graphs the means that were tested in this set of comparisons. (No comparisons between individual soft constraints were carried out, as these failed to be significant in Experiment 4.) This means that three planned comparisons were conducted, i.e., a significance level of $p = .0167$ was used.[16] Note that the Latin square design means that only by-subjects analyses can be carried out here: subjects saw two different lexicalizations for *Ana* (anaphor or pronoun) in a given condition, hence no by-item binding preferences can be computed.

We found a significant difference between the change in binding preference caused

[16] See Footnote 13 on how planned comparisons are handled in this thesis.

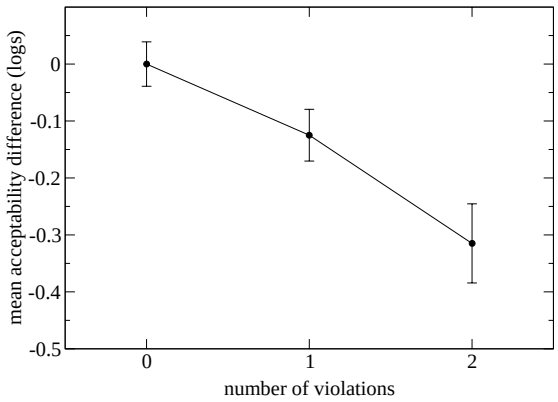

Figure 3.14: Cumulativity of constraint violations (Experiment 5)

by a DEF violation (mean = .0973) and an INT violation (mean = .4031) ($F_1(1,51) = 10.486$, $p = .002$). The difference between a VERB violation (mean = .1585) and an INT violation was also significant ($F_1(1,51) = 6.902$, $p = .011$), as was the difference between a REF violation (mean = .1042) and an INT violation ($F_1(1,51) = 14.938$, $p < .0005$). This means that all soft constraints are ranked higher than INT, which seems to indicate that INT can be classified as a hard constraint.

3.6.5.3. Constraint Interaction

Finally, we wanted to test the hypothesis that constraint interactions are cumulative, in line with the results from Experiment 4. There was only one case of multiple violation in the present experiment, viz., the combined violation of DEF and VERB in the first subexperiment. The associated mean changes in binding preference are graphed in Figure 3.14.

We conducted two planned comparisons, comparing the change in binding preference caused by zero, one, and two violations. The significance level was adjusted according to the Bonferroni method, i.e., we used $p = .025$.[17] As in the first series of planned comparisons, only by-subjects analyses could be carried out. A significant difference was found both between zero violations and one violation (mean = -0.1249) ($F_1(1,51) = 10.953$, $p = .002$), and between one and two violations (mean = -0.3149) ($F_1(1,51) = 12.911$, $p = .001$). This confirms the finding that constraint violations are cumulative (see Experiment 4).

Furthermore, we wanted to test the hypothesis that constraints interact according to the principle of strict domination (see Experiment 4). We conducted a post-hoc test to determine if the change in binding preference caused by the combined violation of DEF and VERB (mean = $-.2964$) was higher than that caused by a single violation of INT (mean = $-.4031$), see Figure 3.13. This difference failed to be significant. (As in Experiment 4, the post-hoc test

[17] See Footnote 13 on how planned comparisons are handled in this thesis.

employed the same significance level as the planned comparisons used to determine constraint ranks, i.e., $p = .0167$.)

This result provides further evidence against strict domination, and for the ganging up of two lower ranked constraints against a higher ranked constraint, already observed in Experiment 4.

3.6.6. Discussion

3.6.6.1. Constraints

This experiment demonstrated that binding in picture NPs is not equally acceptable for anaphors and pronouns. For examples such as (3.37a), we found a main effect of *Ana* (NP type), which shows that pronouns are consistently less acceptable than anaphors. Binding theory, as commonly formalized in various frameworks, expects pronouns to be grammatical in picture NPs, but has to take some extra measures to account for the acceptability of anaphors in the same configurations. For example, Chomsky (1986) introduces the notion of counterfactual coindexation to extend the domain of anaphoric binding in such cases. Pollard and Sag (1994) exempt anaphors in picture NPs from binding theory altogether (as long as there is no referential possessor in the picture NP). They argue that the reference of such anaphors is governed by discourse and processing constraints, which they never explicitly spell out (although they do give a sketch of certain relevant factors). Our results suggest that anaphors should actually be treated as the base case and that it is pronouns that are marginal and exceptional in picture NPs.[18]

We also tested cases where another potential binder intervenes between the pronoun or anaphor and its antecedent. In this case, the acceptability of pronouns and anaphors is not significantly different. Again this contradicts claims in the theoretical literature. When there is an intervening binder, as in (3.42) above, binding theory predicts that only a pronoun should be able to have an antecedent outside the picture NP (as in (3.42b)). We found that the anaphor decreases in acceptability and the pronoun increases in acceptability compared to the case with no intervening binder (see (3.40)). The result is that both forms are equally acceptable, not that the pronoun is acceptable and the anaphor unacceptable (as claimed in the literature). In the control condition, where the binder was inside the NP (see (3.42b)), we found that anaphors are highly acceptable, while pronouns are highly unacceptable, as predicted by binding theory.

We also investigated the factors that influence the exempt status of anaphors. An interaction of *Def* and *Ana* and an interaction of *Verb* and *Ana* was obtained. This demonstrated

[18] One could ask whether our results could be an artifact of the fact that we used linguistically naive speakers, who failed to apply the concept of coreference as intended. Note that the fillers we ran in our experiment were a replication of Gordon and Hendrick's (1997) Experiments 1–4, which tested very basic binding facts (such as the ones in (3.36)). The results we obtained closely matched Gordon and Hendrick's original results, which indicates that our subjects did use the concept of coreference as intended. The replication study is presented in more detail in Chapter 5, Experiment 14.

that the acceptability of pronouns improves if the picture NP is definite or if the matrix verb is a [−EXISTENCE] achievement verb or a [+EXISTENCE] accomplishment verb. However, this improvement in acceptability does not compensate for the general unacceptability of pronoun binding in picture NPs (see Figure 3.10). The verb class effect is line with the claims in theoretical literature (Chomsky 1986; Chomsky and Lasnik 1995; Reinhart and Reuland 1993).

Another finding concerns cases where another potential binder intervenes between the pronoun or anaphor and its antecedent. Here, we observed a reduction in the acceptability of the pronoun if the binder is a quantified NP. This was evidenced by the interaction of *Ref* and *Ana* in our data. In the control condition, where the binder was inside the NP, we failed to find an effect of referentiality, i.e., referential and quantified NPs were equally unacceptable.

These results demonstrate that the constraints that were observed to have an influence on extraction from picture NPs in Experiment 4 (DEF, REF, and VERB) also play a role in binding in picture NPs. This suggests that both phenomena should receive a unified theoretical treatment.

3.6.6.2. Constraint Ranking

The present experiment showed that the constraints DEF, REF, and VERB have a weak influence on the acceptability of binding in picture NPs. The constraint INT (that prevents binding to an anaphor if there is an intervening potential binder), on the other hand, has a strong influence on the acceptability of binding in picture NPs. Its ranking was shown to be significantly higher than that of the other constraints (see also Figure 3.13).

This pattern of results is consistent with the findings obtained in Experiments 4, where we concluded that DEF, REF, and VERB are soft constraints whose violation triggers only small changes in acceptability. The constraint INT, on the other hand, seems to be a hard constraint whose violation leads to a substantial change in acceptability. This claim is supported by the fact that INT outranks all the soft constraints.

3.6.6.3. Constraint Interaction

As for constraint interaction, the present findings confirm the results of Experiment 4, where we provided evidence that constraint violations are cumulative: the combined violation of DEF and VERB leads to an acceptability difference that is significantly higher than that brought about by single violations of these constraints (see Figure 3.13).

Furthermore, we found additional evidence against the strict domination of constraints. The constraint INT was ranked higher than both DEF and VERB. However, a combined violation of DEF and VERB was not significantly different from a single violation of INT. Such a ganging up of constraint violations should be impossible under strict domination; the combination of two lower ranked violations should not compensate for a single violation of a higher

ranked constraint (see again Figure 3.13).

Note this finding shows that two soft constraints (like DEF and VERB) can gang up against a hard constraint (like INT). This allows us to exclude a scenario where hard constraints strictly dominate soft ones, and ganging up effects are restricted to multiple hard violations. Such a scenario could not be included based on the findings of Experiment 4, where ganging up effects were only observed for hard constraints, and multiple soft violations were found to be less serious than a single hard violation.

3.6.7. Conclusions

Following from Experiment 4, the present experiment extended the study of constraint ranking and constraint interaction to a new phenomenon: binding of anaphors and pronouns in picture NPs. The results confirm our classification of constraints: soft constraints cause mild unacceptability when violated, while hard ones lead to serious unacceptability. For binding in picture NPs, we identified definiteness, referentiality, and verb class as soft constraints, while the presence of an intervening potential binder showed a violation pattern characteristic of a hard constraint.

This observation was confirmed by the fact that the hard constraint was found to be ranked significantly higher than all three soft constraints. As far as constraint interaction is concerned, we found evidence for the claim that constraint violations are cumulative. Also, the scope of ganging up effects could be extended to include soft constraints, which constitutes further evidence against strict domination.

3.7. Experiment 6: Effect of Case and Pronominalization on Word Order

Experiments 4 and 5 demonstrated that constraint violations are cumulative. However, this finding was limited to cases of multiple violations of *different* constraints in a given structure. Intuitively, we would expect that this cumulativity effect extends to multiple violations of the *same* constraint. Testing this intuition is the purpose of the present experiment. Note that our results on multiple violations (both of the same constraint and of different constraints) will become important for the theoretical argumentation in Chapter 6.

A second aim of the present experiment is to extend the results on constraint ranking and constraint interaction obtained in Experiments 4 and 5 to a new linguistic phenomenon: word order variation. (We will return to word order extensively in Experiments 10–12, where we will deal with the interaction of word order and context.)

3.7.1. Background

German has a fixed verb order. Subordinate clauses are verb final, while yes/no questions require verb initial order, and declarative main clauses have the verb in second position. In the generative literature, the subordinate clause order is generally considered the basic order from which the main clause and question orders are derived by movement (e.g., Haider 1993). The present experiment (and the follow-up study in Experiment 10) will focus on subordinate clauses (which is also customary in the processing literature on German, e.g., Bader and Meng 1999). Using subordinate clauses avoids potential confounds from topicalization and other phenomena that can occur in verb second clauses.

While verb order is fixed in German, the order of the complements of the verb is variable, and a number of factors have been claimed to influence complement order. These factors include case marking, thematic roles, pronominalization, information structure, intonation, definiteness, and animacy (Choi 1996; Jacobs 1988; Müller 1999; Uszkoreit 1987).

Our approach to word order variation borrows from two existing accounts, Müller (1999) and Uszkoreit (1987). These approaches are interesting in the context of the present thesis because they explicit acknowledge the gradient nature of word order variation, and propose linguistic frameworks that account for gradience. We will test Müller's (1999) and Uszkoreit's (1987) accounts against experimentally collected acceptability judgments (both authors rely on intuitive, informal judgments only).

Uszkoreit (1987) models word order preferences using weighted constraints. In such a setting, linguistic constraints are annotated with a numeric weight that reflects their importance in determining grammaticality (for a similar proposal, see Jacobs 1988). Uszkoreit assumes constraint competition, i.e., not all constraints are necessarily satisfiable in a given linguistic structure. This entails that grammaticality is a gradient notion; the degree of grammaticality of a linguistic structure is computed as the sum of the of the weights of the constraint violations the structure incurs.

Uszkoreit (1987) proposes the following constraints on word order in German (we omit constraints that are not relevant to the present study):

(3.43) a. $\mathrm{V}[+\mathrm{MC}] \prec \mathrm{X}$
b. $\mathrm{X} \prec \mathrm{V}[-\mathrm{MC}]$
c. $[+\mathrm{NOM}] \prec [+\mathrm{DAT}]$
d. $[+\mathrm{NOM}] \prec [+\mathrm{ACC}]$
e. $[+\mathrm{DAT}] \prec [+\mathrm{ACC}]$
f. $[-\mathrm{FOCUS}] \prec [+\mathrm{FOCUS}]$
g. $[+\mathrm{PRO}] \prec [-\mathrm{PRO}]$ (Uszkoreit 1987: 114)

These constraints are constituent order constraints, with "$\prec$" denoting linear precedence. The constraint (3.43a) relies on the feature MC (main clause) to specify verb order; if this feature

is positive (i.e., in a main clause), then the verb has to precede any other constituent, resulting in verb initial word order. In a subordinate clause (marked [−MC]), on the other hand, all other constituents have to precede the verb, as specified by constraint (3.43b), resulting in verb final order. The constraints (3.43c) and (3.43d) require that nominative NPs precede dative and accusative NPs. The information structural requirement (3.43f) specifies that ground constituents (marked [−FOCUS]) precede focused constituents. Finally, the constraint (3.43g) requires pronouns to precede full NPs.

Uszkoreit does not provide ranks or weights for the constraints in (3.43). Intuitively, however, we expect a violation of verb order to lead to serious unacceptability, i.e., constraint (3.43b) should receive a higher weight than the other constraints. Also Pechmann, Uszkoreit, Engelkamp, and Zerbst (1994), who use a slightly modified version of the constraint set in (3.43), assume that [+NOM] ≺ [+DAT] and [+NOM] ≺ [+ACC] are stronger than the constraints on verb order in (a) and (b).

An alternative to Uszkoreit's (1987) approach has been proposed by Müller (1999) based on Optimality Theory. Standard Optimality Theory (Prince and Smolensky 1993, 1997) assumes a binary notion of grammaticality; a linguistic structure is either optimal (and thus grammatical) or suboptimal (and thus ungrammatical). However, OT can be extended to model gradient grammaticality; Müller (1999) puts forward a modified version of OT based on the distinction between grammaticality (manifested in binary judgments) and markedness (associated with word order preferences). Grammaticality is handled in terms of conventional OT-style constraint competition. This competition can yield several grammatical candidates, among which further competition takes place based on markedness constraints. The markedness competition then induces a preference order on the candidates that predicts their relative acceptability. (Note that the grammaticality/markedness dichotomy is reminiscent of the distinction of hard and soft constraints proposed in this thesis.)

In Müller's account, the constraints on pronoun order belong to the realm of grammaticality, while the constraints on case order and focus-ground order (among others) belong to the realm of markedness. We omit technical details and only state constraints relevant to the present data set:

(3.44) a. NOM: [+NOM] ≺ [−NOM]
b. FOC: [−FOCUS] ≺ [+FOCUS]
c. DAT: [+DAT] ≺ [+ACC] (Müller 1999: 795)

Note that the constraint NOM has the same effect as the constraints (3.43c) and (3.43d) postulated by Uszkoreit. Also the constraint FOC is the same as Uszkoreit's (3.43f).

In contrast to Uszkoreit, Müller postulates an explicit constraint ranking ("≫" stands for "is ranked higher than"):

(3.45) NOM ≫ FOC ≫ DAT

In addition to the markedness constraints in (3.44), a set of grammaticality constraints is postulated (omitted here). These constraints deal with pronoun order and ensure that pronouns occur at the left periphery of the clause. All candidates that fail to meet this requirement are predicted to be (categorically) ungrammatical. In contrast to Uszkoreit, Müller does not include constraints on verb order; however it seems safe to assume that such constraints would be grammaticality constraints in Müller's system.

For the purpose of this thesis, we will assume a set of constraints that is based on the constraints assumed by Uszkoreit and Müller:

(3.46) **Constraints on Word Order and Information Structure**

a. VERBINITIAL: V[+MC] $\prec$ X
b. VERBFINAL: X $\prec$ V[−MC]
c. PROALIGN: [+PRO] $\prec$ [−PRO]
d. NOMALIGN: [+NOM] $\prec$ [−NOM]
e. DATALIGN: [+DAT] $\prec$ [+ACC]
f. GROUNDALIGN: [−FOCUS] constituents have to be peripheral.

The present experiment will test the validity of the constraints PROALIGN, NOMALIGN, and DATALIGN, while the follow-up Experiment 10 will deal with the additional constraint VERBFINAL and GROUNDALIGN.

Note that we have adopted a formulation of GROUNDALIGN that differs from the one proposed by Uszkoreit and Müller. This formulation requires that ground constituents (marked [−FOCUS]) are at the peripheral, i.e., occur sentence initially or sentence finally. This constraint makes the same predictions as [−FOCUS] $\prec$ [+FOCUS] for verb final orders, but makes different predictions for verb initial and verb medial orders. This will become important in Experiments 11 and 12, where we will investigate word order preferences in Greek based on the constraint GROUNDALIGN. These experiments will also include new constraints on clitic doubling and accent placement. The verb ordering restriction VERBINITIAL will be used in the modeling studies in Chapter 7 (but has not been investigated experimentally).

Previous judgment studies on word order in German were reported by Pechmann et al. (1994) and Scheepers (1997). Pechmann et al. (1994) based their investigation on Uszkoreit's (1987) set of constraints and were able to largely confirm his predictions, using both judgments and a number of processing and production tasks. Scheepers (1997) focused on the interaction of syntactic constraints (such as nominative precedes accusative) with thematic constraints (such as agent precedes patient) and concluded that syntactic constraints are stronger than thematic ones (again based on evidence from both judgments and processing tasks). Neither of these two studies dealt with the effects of pronominalization on word order preferences (which is the focus of the present experiment) or with context effects (which will be addressed in Experiment 10; see also Meng, Bader, and Bayer 1999).

3.7.2. Introduction

The aim of this experiment is to establish how multiple violations of the constraints PROALIGN, NOMALIGN, and DATALIGN influence the acceptability of a given structure.

The experiment uses ditransitive verbs such as *vorschlagen* "propose" that can take three animate NPs as arguments. All possible permutations of these three NPs are tested, see the example in (3.47).[19] Our notation for word orders uses "V" for verb, "S" for subject, and "O" and "I" for direct and indirect object, respectively. Subscript "pro" is used to indicate that the NP is pronominalized.

(3.47) a. **SIOV:** Ich glaube, dass der Produzent dem Regisseur den Schauspieler
I believe that the producer-NOM the director-DAT the actor-ACC
vorschlägt.
proposes
"I believe that the producer will propose the actor to the director."

b. **SOIV:** Ich glaube, dass der Produzent den Schauspieler dem Regisseur vorschlägt.

c. **ISOV:** Ich glaube, dass dem Regisseur der Produzent den Schauspieler vorschlägt.

d. **IOSV:** Ich glaube, dass dem Regisseur den Schauspieler der Produzent vorschlägt.

e. **OSIV:** Ich glaube, dass den Schauspieler der Produzent dem Regisseur vorschlägt.

f. **OISV:** Ich glaube, dass den Schauspieler dem Regisseur der Produzent vorschlägt.

These orders allow us to test the effect of violations of NOMALIGN and DATALIGN. The order SIOV does not violate any constraints. SOIV violates DATALIGN once, as the accusative NP precedes the dative NP. ISOV violates NOMALIGN once as there is one non-nominative NP that precedes the nominative NP, while in IOSV, two non-nominative NPs precede the nominative NP, hence this structure incurs a double violation of NOMALIGN.

The examples in (3.47) also allow us to test combined violations of NOMALIGN and DATALIGN. OSIV violates both NOMALIGN and DATALIGN, as the accusative (non-nominative) NP precedes the nominative NP, and the accusative NP also precedes the dative NP. Finally, OISV violates NOMALIGN twice, as both the accusative and the dative NP precede the nominative NP, and also incurs a violation of DATALIGN, as the accusative NP precedes the dative NP

The second part of the experiment deals with the predictions of the constraint PROALIGN. We use the same six orders as in (3.47), but now one of the NPs is realized as

[19] Note that only masculine NPs were used, as these are unambiguous in their case marking, both as full NPs and as pronouns (while the case morphology of feminine and neuter NPs exhibits syncretism).

a pronoun. The position of the pronominalized NP varies; either the first, second, or third NP is realized as a pronoun. This is illustrated in the following example for the order SIOV.

(3.48) a. **S_{pro}IOV:** Ich glaube, dass er dem Regisseur den Schauspieler
I believe that he-NOM the director-DAT the actor-ACC
vorschlägt.
proposes
"I believe that he will propose the actor to the director."

b. **SI$_{pro}$OV:** Ich glaube, dass der Produzent ihm den Schauspieler
I believe that the producer-NOM him-DAT the actor-ACC
vorschlägt.
proposes
"I believe that the producer will propose the actor to him."

c. **SIO$_{pro}$V:** Ich glaube, dass der Produzent dem Regisseur ihn
I believe that the producer-NOM the director-DAT him-DAT
vorschlägt.
proposes
"I believe that the producer will propose him to the director."

These sentences incur zero to two violations of PROALIGN. (3.48a) violates PROALIGN zero times, since there is no full NP that precedes the pronoun. (3.48b) incurs one violation, as one full NP precedes the pronoun. In (3.48c), there are two full NPs preceding the pronoun, hence this sentence incurs two violations of PROALIGN.

3.7.3. Predictions

3.7.3.1. Constraints

In line with the observations by Uszkoreit (1987), Müller (1999) (among others), we expect that a violation of the constraints NOMALIGN, DATALIGN, and PROALIGN will lead to a significant reduction in the acceptability of a given word order. This means we predict main effects of the corresponding experimental factors *Nom*, *Dat*, and *Pro*.

3.7.3.2. Constraint Ranking

Recall that we adopted an operational definition of constraint ranking based on the degree of unacceptability caused by a given constraint violation (see Section 3.1.2). This definition can be used to assess the ranking of the constraints dealt with in this experiment: we will compare the degree of unacceptability caused by single violations of NOMALIGN, DATALIGN, and PROALIGN. Differences in unacceptability will indicate differences in constraints ranking, and can be tested using planned comparisons.

Predictions with respect to the ranking of NOMALIGN, DATALIGN, and PROALIGN can be arrived at based on Müller's (1999) OT analysis. He assumes the ranking NOM ≫ DAT, where his constraints NOM and DAT correspond to our constraints NOMALIGN and DATALIGN. Furthermore, he stipulates that the order of pronouns is governed by grammaticality constraints (hard constraints in our terminology). This means that his account predicts that violations of PROALIGN are more serious than violations of NOMALIGN, because NOMALIGN is a markedness constraint in Müller's (1999) account (a soft constraint in our terminology).

3.7.3.3. Constraint Interaction

In Experiments 4 and 5, we established that multiple constraint violations have an cumulative effect on acceptability. This effect was found for structures that incur multiple violations of different constraints. The purpose of the present experiment is to determine how multiple constraint violations of the same constraint influence acceptability.

This can be established with respect to the constraints NOMALIGN and PROALIGN, for which we include single and double violations in the stimulus set. If we assume that the cumulativity of constraint violations generalizes to multiple violations of the same constraint, then we predict that double violations of NOMALIGN and PROALIGN trigger a higher degree of unacceptability than single violations.

If we find that multiple violations of the same constraint are cumulative, then we also expect an overall cumulativity effect. To test this hypothesis, we can conduct an analysis based on the overall number of violations in a given stimulus, irrespective of whether they are violations of the same constraint or of different constraint. This allows us to investigate structures with up to three violations (for non-pronominalized stimuli), or with up to five violations (for pronominalized stimuli). As in previous experiments, these predictions regarding cumulativity will be tested using a series of planned comparisons.

3.7.4. Method

3.7.4.1. Subjects

Twenty-seven native Speakers of German from the same population as in Experiment 1 participated in the experiment.

The data of two subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 25 subjects for analysis. Of these, 17 subjects were male, eight female; 22 subjects were right-handed, three left-handed. The age of the subjects ranged from 16 to 41 years, the mean was 27.3 years.

3.7.4.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials The experiment used two subdesigns. The first subdesign dealt with non-pronominalized noun phrases and crossed the factors *Nom* and *Dat*. The factor *Nom* had three levels, specifying the number of violations of the constraint NOMALIGN (once, twice, three times). The factor *Dat* had only two levels: either the constraint *DatAlign* was violated or not. By crossing the factors *Nom* and *Dat*, we arrive at six cells, which correspond to the six word orders given in (3.47). Eight lexicalizations were used for each of the cells, which resulted in a total of 48 stimuli.

The second subdesign investigated the same word orders as the first subdesign, but this time, one of the three NPs was pronominalized. This was realized by the additional factor *Pro*, which had three levels, specifying the number of violations of the constraint PROALIGN (once, twice, three times). This yielded $Nom \times Dat \times Pro = 3 \times 2 \times 3 = 18$ cells in total. Example stimuli are given in (3.48). Each cell was realized by the same eight lexicalizations as in the first subdesign, resulting in 144 stimuli.

A set of 24 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

To control for possible effects from lexical frequency, the lexicalizations for subject, direct object, indirect object, and verb were matched for frequency. Frequency counts for the verbs and the head nouns were obtained from a lemmatized version of the Frankfurter Rundschau corpus (40 million words of newspaper text) and the average frequencies were computed for subject, direct object, indirect object, and verb lexicalizations. An ANOVA confirmed that these average frequencies were not significantly different from each other.

3.7.4.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions, Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

Eight test sets were used: each test set contained one lexicalization for each of the six cells in the first subdesign, and one lexicalization for each of the 18 cells in the second subdesign, i.e., a total of 24 items. Lexicalizations were assigned to test sets using a Latin square covering the full set of items.

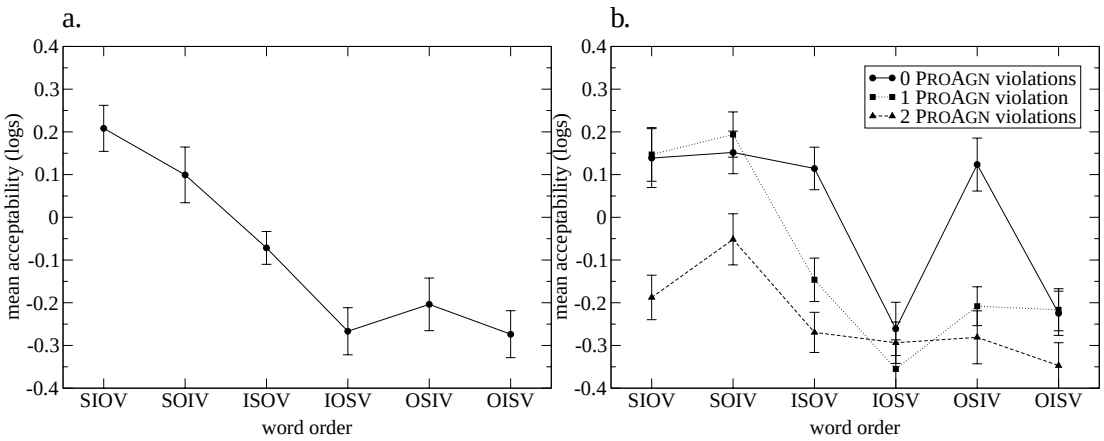

Figure 3.15: Mean judgments for each word order for (a) the non-pronominalized and (b) the pronominalized condition (Experiment 6)

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 48 test items: 24 experimental items and 24 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to one of the test sets.

3.7.5. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

3.7.5.1. Constraints

The mean acceptability ratings for each word order for the first subexperiment are displayed in Figure 3.15a. An ANOVA revealed a main effect of *Nom* that was highly significant ($F_1(2,48) = 30.197$, $p < .0005$; $F_2(2,14) = 23.125$, $p < .0005$), while the main effect of *Dat* was significant by subjects and marginal by items ($F_1(1,24) = 5.710$, $p = .025$; $F_2(1,7) = 3.563$, $p = .101$). The interaction of *Nom* and *Dat* failed to be significant.

Figure 3.15b graphs mean acceptability for each word order for the second subexperiment, which included pronominalized NPs. The ANOVA again showed a highly significant main effect of *Nom* ($F_1(2,48) = 43.410$, $p < .0005$; $F_2(2,14) = 25.236$, $p < .0005$). A highly significant main effect of *Pro* was also obtained ($F_1(2,48) = 31.945$, $p < .0005$; $F_2(2,14) = 24.058$, $p < .0005$), while the main effect of *Dat* was not significant. Furthermore, an interaction of *Nom* and *Pro* was present ($F_1(2,48) = 10.864$, $p < .0005$; $F_2(2,14) = 24.058$, $p < .0005$). All other interactions were not significant, with the exception of the three-way interaction of *Nom*, *Dat*, and *Pro*, which was significant by subjects only ($F_1(4,96) = 2.668$, $p = .037$; $F_2(4,28) = 1.073$, $p = .388$).

3.7.5.2. Constraint Ranking

As in Experiments 4 and 5, a separate analysis was conducted to determine constraint rankings. The first subexperiment allows us to compare NOMALIGN and DATALIGN violations. We carried out a planned comparison and found that the unacceptability caused by a single NOMALIGN violation (order ISOV, mean $= -.0779$) was higher than the unacceptability caused by a single DATALIGN violation (order SOIV, mean $= .0963$); this difference was significant by subjects and marginal by items ($F_1(1,24) = 5.300$, $p = .030$; $F_2(1,7) = 3.809$, $p = .092$). (The mean acceptability of these orders is graphed in Figure 3.15a.)

The second subexperiment allowed us to compare NOMALIGN and PROALIGN violations using a planned comparison of the degree of unacceptability caused by single constraint violation. No significant difference was found between the unacceptability caused by a NOMALIGN (order I_{pro}SOV, mean $= .1144$) and a PROALIGN violation (order SI_{pro}OV, mean $= .1471$). Note that no comparisons involving DATALIGN were carried out, as DATALIGN failed to have a significant effect on acceptability in the second subexperiment. (The mean acceptability of these orders is graphed in Figure 3.15b.)

3.7.5.3. Constraint Interaction

Figure 3.16a graphs the mean acceptability for zero, one, and two violations of NOMALIGN, and for zero or one violation of DATALIGN for the non-pronominalized stimuli. To investigate if multiple violations of NOMALIGN are cumulative, we conducted a post-hoc Tukey test on the main effect of *Nom* that was found in the first subexperiment. Stimuli with one violation of NOMALIGN (mean $= -.1377$) were less acceptable than stimuli with zero violations of NOMALIGN (mean $= .1538$) ($\alpha < .01$). Stimuli with two violations (mean $= -.2701$) were in turn less acceptable than stimuli with one violation (by subjects only, $\alpha < .05$).

Figure 3.16b graphs the multiple constraint violations for the second subexperiment (pronominalized stimuli) for the constraints NOMALIGN, DATALIGN, and PROALIGN. Again, a post-hoc test was conducted based on the main effect of *Nom* that was found for the second subexperiment. The results confirmed the findings from the first subexperiment: one violation of NOMALIGN (mean $= -.1111$) was less acceptable than zero violations (mean $= .0654$) ($\alpha < .01$), while two violations (mean $= -.2830$) were less acceptable than one violation ($\alpha < .01$). A further post-hoc test on the main effect of *Pro* demonstrated the cumulativity of PROALIGN violations: one violation of PROALIGN (mean $= -.0975$) was less acceptable than zero violations of PROALIGN (mean $= .0071$) ($\alpha < .01$), while two violations (mean $= -.2384$) were less acceptable than one violation ($\alpha < .01$).

Furthermore, a series of planned comparisons was carried out to establish the presence of a general cumulativity effect for multiple violations of either the same or different constraints. In the second subexperiment, each stimulus incurred up to four violations of

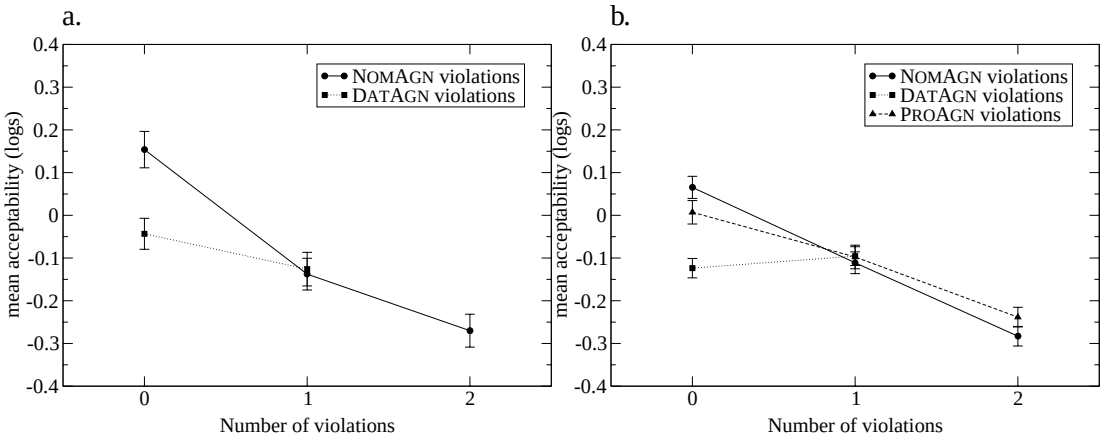

Figure 3.16: Cumulativity of violations for NOMALIGN, DATALIGN, and PROALIGN in (a) the non-pronominalized and (b) the pronominalized condition (Experiment 6)

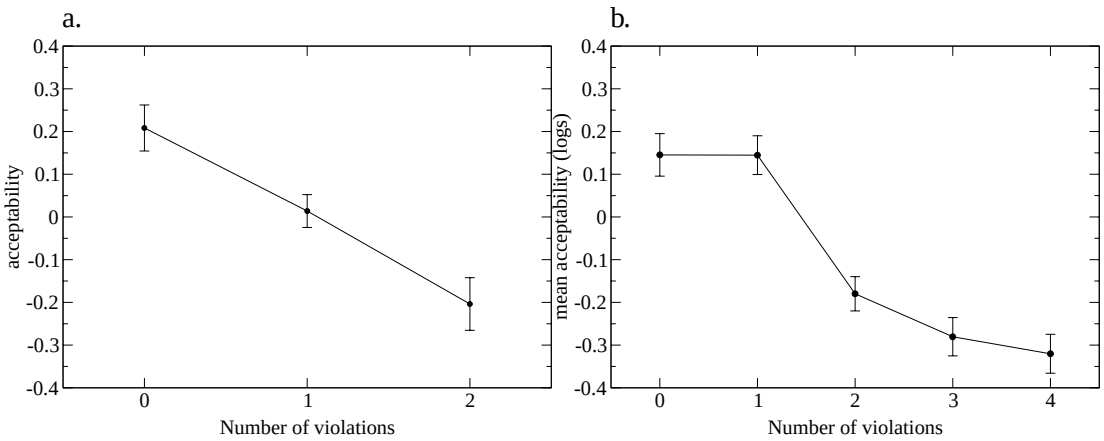

Figure 3.17: Cumulativity of constraint violations for (a) the non-pronominalized and (b) the pronominalized condition (Experiment 6)

NOMALIGN or PROALIGN. (Violations of DATALIGN were not taken into account since no main effect of *Dat* was found in the pronominalized condition.) As in Experiments 4 and 5, we computed the mean acceptability for the stimuli with zero violations (one sentence type), one violation (two sentence types), two violations (three sentence types), three violations (two sentence types), and four violations (one sentence type). The resulting acceptability scores are graphed in Figure 3.17b.

Four planned comparisons were carried out, hence we set the significance level at $p = .0125$. We failed to find a difference between zero violations (mean $= .1453$) and a single violation (mean $= .1447$). However, the difference between a single violation and a double violation (mean $= -.1799$) was significant ($F_1(1,24) = 51.745$, $p < .0005$; $F_2(1,7) = 50.044$, $p < .0005$). Also, there was a difference between a double violation and a triple violation

(mean $= -.2804$), significant by subjects only ($F_1(1,24) = 7.143$, $p = .013$; $F_2(1,7) = 4.208$, $p = .079$). The difference between a triple violation and a quadruple violation (mean $= -.3203$), on the other hand, failed to reach significance.

A test for cumulativity effects was also carried out for the first subexperiment. Here we included stimuli with zero violations, a single violation of either NOMALIGN or DATALIGN, and a double violation of both NOMALIGN and DATALIGN. Note that stimuli with double violations of NOMALIGN were not included as this would create an asymmetrical design, given that NOMALIGN was ranked higher than DATALIGN (recall that no double violations of DATALIGN were tested). Two planned comparisons were carried out, hence we set the significance level at $p = .025$. We found that a single violation (mean $= .0139$) was significantly more serious than zero violations (mean $= .2083$) ($F_1(1,24) = 15.700$, $p = .001$; $F_2(1,7) = 12.672$, $p = .009$). Also, a double violation (mean $= -.2038$) was more serious than a single violation (significant by subjects and marginal by items, $F_1(1,24) = 16.041$, $p = .001$; $F_2(1,7) = 6.001$, $p = .044$).

3.7.6. Discussion

3.7.6.1. Constraints

We found significant main effects of the factors *Nom*, *Dat*, and *Pro*, which confirms our prediction that a violation of either of the three constraints NOMALIGN, DATALIGN, and PROALIGN leads to a significant reduction in acceptability.

However, the effect of *Dat* was only present in the first subexperiment; for pronominalized orders (subexperiment two), the relative order of dative and accusative NPs does not seem to influence acceptability (see also Figure 3.15b). This is an interesting finding that is not predicted by the theoretical approaches our constraints are based (Müller 1999; Uszkoreit 1987).

Another finding concerns the interaction of *Nom* and *Pro* that was present in the second subexperiment. The meaning of this interaction becomes clear from Figure 3.15b: the impact of NOMALIGN depends on the whether or not violations of PROALIGN are also present. For orders with a single PROALIGN violation, we find a pattern that corresponds to the one found for non-pronominalized orders, modulo the effect of DATALIGN, which was not present in the second subexperiment (compare Figures 3.15a and 3.15b). For double violations of PROALIGN, we find essentially the same pattern as in the one violation condition (though overall acceptability is low, which is of course due to the double violation of PROALIGN). However, the pattern for orders with zero PROALIGN violations deviates from that found in the non-pronominalized condition. A single violation of NOMALIGN does not seem to have an effect, while a double violation of NOMALIGN causes a sharp drop in acceptability (compare the acceptability of I_{pro}SOV or O_{pro}SIV with that of I_{pro}OSV or O_{pro}ISV).

The same phenomenon can also be observed for single violations of PROALIGN; these did not cause a drop in acceptability, while double violations are severely unacceptable (compare the acceptability of S_{pro}IOV or S_{pro}OIV with that of SI$_{pro}$OV or SO$_{pro}$IV). Once another violation of PROALIGN or a violation of NOMALIGN is present, an additional violation of PROALIGN has the expected effect. We do not have an explanation for this finding, which is not reported in the literature.

However, our findings on non-pronominalized word orders are consistent with those reported in the experimental literature. Pechmann et al. (1994) found the following acceptability ranking for non-pronominalized orders: SIOV $>$ SOIV $>$ ISOV $>$ OSIV $>$ IOSV.[20] This ranking matches the one obtained in the first subexperiment (see Figure 3.15), even though Pechmann et al.'s (1994) study differed from the present one in using inanimate accusative NPs. We conclude that animacy fails to have an effect on the order preferences (or that its effect is weak compared to the effects of NOMALIGN and DATALIGN).

3.7.6.2. Constraint Ranking

We established that the constraint NOMALIGN is ranked higher than DATALIGN. This ranking is compatible with Müller's (1999) account: he assumes a ranking NOM $\gg$ DAT, where his constraints NOM and DAT correspond to our constraints NOMALIGN and DATALIGN.

On the other hand, we failed to find a difference in ranking between the constraints NOMALIGN and PROALIGN. This conflicts with Müller's (1999) analysis, which assumes that the position of pronouns is governed by hard constraints, while the position of nominative NPs is governed by soft constraints. In contrast, our results indicate that the order of both pronouns and nominative NPs is governed by soft constraints (of equal ranking). This finding will be confirmed in Experiment 10, where will provide further evidence for the status of NOMALIGN and PROALIGN as soft constraints by demonstrating that these constraints are subject to context effects. We will also compare these constraints to a genuine hard constraint, viz., the constraint that regulates verb order in German.

On a more general level, this result points to the importance of using experimental methods to obtain gradient judgments data—Müller's (1999) analysis is based on informal, intuitive judgments only.

3.7.6.3. Constraint Interaction

The experimental findings provided clear evidence for the generality of the cumulativity effect. We demonstrated that two violations of a given constraint trigger a higher degree of unacceptability than a single violation of the same constraint. This was demonstrated for the constraints

[20] We will use "$>$" to denote "is more acceptable than".

NOMALIGN and PROALIGN, for which we included single and double violations in the stimulus set.

Given that multiple violations of the same constraint are cumulative, we also expected an overall cumulativity effect. This was confirmed in a series of planned comparisons where we counted the number of violation incurred by a stimulus, irrespective of whether they were violations of the same constraint or of different constraints. We found evidence for the cumulativity for up to four violations of NOMALIGN, DATALIGN, and PROALIGN.

An unexpected effect occurred in the pronominalized condition, where we found that stimuli with one constraint violation were as acceptable as stimuli with no violations (see Figure 3.13). The reason for this was already discussed in Section 3.7.5.1: it seems that the effect of NOMALIGN and PROALIGN only becomes visible once another violation of either NOMALIGN or PROALIGN is present (see Figure 3.17b).

3.7.7. Conclusions

Investigating word order as a new phenomenon, the present experiment demonstrated that multiple violations of the same constraint are cumulative. This extends the results on the cumulativity of violations of the different constraints obtained in Experiments 4 and 5.

Taken together, Experiments 4–6 suggest that cumulativity is a general property of constraint violations. An cumulativity effect could be demonstrated for both soft and hard violations, and for multiple violations of the same constraint and of different constraints. These findings are not readily compatible with an OT-style model of constraint interaction, where the optimality of a structure is determined primarily based on the rank of the constraints it violates, rather than based on the number of constraint violations. We will return to this in our theoretical discussion in Chapter 6.

The present experiment also provided more evidence for the ranking of constraints; we showed that the constraint NOMALIGN is ranked higher than the constraint DATALIGN, based on the degree of unacceptability caused by violations of NOMALIGN and DATALIGN, respectively. NOMALIGN and DATALIGN are soft constraints (this was not demonstrated in the present experiment, but will become clear in Experiment 10 in the next chapter). Hence this result about constraint ranking complements the finding of Experiment 4, where we only found evidence for the ranking of hard constraints.

3.8. Conclusions

As detailed in Section 3.1, the present chapter had a double purpose. Firstly, it investigated a set of linguistic constraints by presenting experimental results on four syntactic phenomena (unaccusativity, extraction, binding, and word order). The results show that the use of gradient acceptability judgments (collected experimentally) can contribute to clarifying the empirical

status of these constraints, thus allowing us to settle data disputes in linguistic theory. The underlying assumption is that such data disputes are the results of the informal data collection techniques employed in theoretical linguistics, which are not well-suited to investigate the behavior of gradient linguistic data.

The second purpose of the experiments in the present chapter was to provide initial evidence concerning a number of general properties of gradient linguistic data. These properties concern the classification of constraints into types, and the ranking and interaction of constraints.

Experiments 1–3 focussed on constraint types. The results led to a distinction of hard and soft constraints, based on the following set of properties:

- **Gradience** Soft constraint violations are associated with mild unacceptability, while hard violations trigger serious unacceptability. This explains why hard constraints are intuitively associated with binary acceptability judgments, while soft ones are associated with degrees of acceptability.

- **Crosslinguistic Variation** The effects of hard constraints are crosslinguistically stable (we take dialect variation as an instance of crosslinguistic variation). Soft constraints, on the other hand, may exhibit crosslinguistic effects.

Some of these results were preliminary, and require further experimental support. For example, our study of crosslinguistic effects was limited to a comparison of two dialects of German. We will return to this issue in Experiments 10–12, where we present the results of a crosslinguistic study of gradience in word order phenomena.

Experiments 4–6 focussed on constraint ranking and constraint interaction. Our investigation of these concepts was guided by a set of operational definitions (see Sections 3.1.2 and 3.1.4); the ranking of a constraint was defined as the degree of unacceptability caused by its violation. Our experimental results demonstrated the following properties:

- **Ranking** Both soft and hard constraints are ranked, i.e., individual constraints may differ in the degree of unacceptability they incur when violated. Evidence for the ranking of hard constraints was provided in Experiment 4, where we investigated extraction; the ranking of soft constraints was demonstrated in Experiment 6, where we dealt with word order variation.

- **Cumulativity** Constraint violations are cumulative, i.e., the unacceptability of a structure increases with the number of constraints it violates. It was shown that this is an effect of considerable robustness and generality; it applies to both soft and hard violations, and to multiple violations of the same constraint and of different constraints.

- **Strict Domination** We found evidence for the ganging up of constraint violations, both for hard and for soft constraints (in Experiments 4 and 5). Ganging up effects are

not compatible with OT-style strict domination, but are expected under the assumption that constraint violations are cumulative.

In the next chapter, we will present another set of experiments which is designed to broaden the support for the hypotheses regarding gradient data that were formed in the present chapter. We will provide more data on extraction and word order, and we will deal with a further linguistic phenomenon, gapping.

The next chapter will provide additional support for the cumulativity effect and for constraint ranking. It will also broaden our investigation of the relationship between crosslinguistic variation and gradience. However, the main purpose of the next chapter will be to investigate the relationship between gradience and linguistic context. We will put forward the hypothesis that soft constraints are subject to context effects, while hard constraints are immune to contextual variation. It will be argued that context effects can therefore serve as a diagnostic for the soft/hard distinction. We will present data on context effects on linguistic judgments for three phenomena: gapping (Experiments 7 and 8), extraction (Experiment 9), and word order (Experiments 10–12).

Chapter 4

Gradient Grammaticality in Context

Chapter 3 presented a preliminary investigation into the nature of gradient linguistic judgments, and identified a set of general properties shared by all the syntactic phenomena that were investigated. We found that constraints are ranked, that constraint violations are cumulative, and that lower ranked constraint violations can gang up against higher ranked ones. In this chapter, we report a series of experiments on gapping, extraction, and word order that confirm these basic observations.

Another result of Chapter 3 was the hypothesis that there are two types of constraints that exhibit distinct behavior with respect to gradient judgments. Soft constraints lead to mild unacceptability when violated, while violations of hard constraints trigger serious unacceptability. This chapter will supply additional evidence for the hard/soft dichotomy. We will show that context effects are a powerful diagnostic of constraint type: the experimental findings suggest that soft constraints are subject to context effects, while hard constraints are immune to contextual variation.

In Chapter 3, we hypothesized that the effects of hard constraints are crosslinguistically stable, while soft constraint effects are subject to crosslinguistic variation. This chapter provides further evidence for this hypothesis based on a crosslinguistic investigation of word order preferences.

4.1. Introduction

In Chapter 3, the main focus of our investigation of gradience in grammar was on constraint ranking, constraint types, and constraint interaction. This chapter continues to provide evidence with regard to constraint ranking and constraint interaction; however, its main focus is the hypothesis that two types of linguistics constraints can be distinguished, soft and hard, based on their behavior with respect to gradient acceptability. The experiments reported in the present chapter will explore his hypothesis by demonstrating that crosslinguistic variation and context

effects can serve as diagnostics for the hard/soft dichotomy.

4.1.1. Context Effects

Throughout this thesis the term "context" will be used to refer to the linguistic context of a sentence, i.e., to the type of context that is involved in intersentential grammatical phenomena such as Information Structure (Experiments 7, 8, 10–12) or reference and presupposition (Experiment 9). We will not deal with effects from the extra-linguistic context of a sentence, which are well-attested for linguistic judgments (see Section 2.4 for an overview).

The linguistic context of a sentence can be manipulated by prefixing the target sentence with another sentence, the context sentence. This context sentence can either be declarative or a question (the latter is common for information structural phenomena). Using this approach, we are able to test for context effects in the target sentence by manipulating the linguistic properties of the context sentence. It is important to compare the results of such manipulations to a control condition. This is standard practice in the sentence processing literature (see, e.g., Altmann and Steedman 1988). All experiments reported in the present chapter use a double control condition: a neutral context, i.e., a context that is maximally uninformative, and a null context where the target sentence is presented in isolation.

The focus of this chapter is not on context effects as such, but rather on the interaction of constraint violations with context. To conceptualize this interaction, we will distinguish two kinds of constraints: context-independent constraints and context-dependent constraints. A constraint is *context-independent* if it is immune to context effects, i.e., if its violation causes the same degree of unacceptability in all contexts. A constraint is *context-dependent* if the degree of unacceptability triggered by its violation varies from context to context. An extreme example of a context-dependent constraint is one for which the effect of its violation disappears completely in certain contexts, i.e., the violation triggers no increase in unacceptability. For other constraints we might find that the effect of a violation is less serious in a certain context, i.e., the violation leads to a lower degree of unacceptability compared to other contexts.

The hypothesis that will be advanced in the present chapter is that soft constraints are context-dependent, while hard constraints are context-independent. If this hypothesis is correct, then context effects can serve as a diagnostic for the type of a constraint, i.e., by checking for context effects we can determine if a given constraint is hard or soft.

4.1.2. Crosslinguistic Effects

Experiments 1–3 presented an initial investigation of crosslinguistic variation in gradient data. This investigation led to the conclusion that certain verb classes (peripheral verb classes) show an auxiliary selection behavior that varies from language to language (or from dialect to dialect), while other classes (core verb classes) show the same auxiliary selection behavior in all

languages (Sorace 2000).

This finding can be explained under the assumption that class membership is governed by a set of constraints, some of which are hard constraints, while others are soft ones. The hard constraints determine the membership in core verb classes, while the soft ones regulate the membership in peripheral verb classes. In the preceding chapter, we demonstrated that hard constraints lead to serious unacceptability when violated, while soft constraint violations induce only mild deviance. This explains why core verbs show a strong preference for one auxiliary, while the auxiliary selection preferences of peripheral verbs are gradient.

Under an optimality theoretic approach, crosslinguistic variation is accounted for by the re-ranking of constraints (for an overview of Optimality Theory, see Section 2.6). In the case of auxiliary selection, this means that the crosslinguistic differences in the auxiliary preference of peripheral verbs are due to crosslinguistic differences in the ranking of the soft constraints that govern class membership for peripheral verbs. For instance, the controlled motion class is a peripheral class that selects *sein* in northern dialects, while southern dialects allow both *haben* and *sein* (see Experiment 1). This means that in northern dialects, the constraint that disallows *haben* for controlled motion verbs is ranked higher than the constraint that disallows *sein*, resulting in an overall *sein* preference. In southern dialects, on the other hand, both constraints are ranked equally, resulting in equal acceptability for both auxiliaries. For core verb classes, however, no such crosslinguistic variation in the constraint ranks is predicted. Change of location verbs, for instance, select *sein* in both dialects, which means that the constraint that disallows *haben* for this class has the same rank in both dialects.

In this setting, constraint re-ranking only affects soft constraints, while hard constraints have the same rank across languages. While this might be the case for auxiliary selection, it does not seems to generalize to other syntactic phenomena. In fact, most of the optimality theoretic research on syntax (see the papers in Barbosa, Fox, Hagstrom, McGinnis, and Pesetsky 1998 as an example) accounts for crosslinguistic variation via the re-ranking of hard constraints. (By hard constraints we mean constraints that induce clear-cut, binary acceptability judgments. In the OT literature, however, the term hard constraint is sometimes used to refer to inviolable constraints.)

We will therefore assume that the re-ranking of hard constraints is possible, and that auxiliary selection (where the ranking of hard constraint seems to be fixed) is just a special case. Under this view, crosslinguistic re-ranking is a general property of linguistic constraints, both hard and soft. However, a crucial difference between soft and hard constraints remains, even if we assume that the two constraint types both allow re-ranking. Recall that the results of Experiments 1–3 showed that core verbs are core across languages, i.e., they exhibit a binary auxiliary selection pattern in all languages (or dialects). On the other hand, peripheral verbs are peripheral across languages, i.e., they show gradient auxiliary selection, and are subject to telicity and animacy effects. There seems to be no cases of verb classes that are core in one

language, but peripheral in another.

This observation leads to the hypothesis that constraint re-ranking can not cross type boundaries. In other words, hard constraints will be hard in all languages (but may vary in their ranking from language to language). Correspondingly, soft constraints are crosslinguistically soft (but the ranking can vary crosslinguistically). We will test this hypothesis in the our crosslinguistic investigation of word order in Experiments 10–12. The crosslingusitic behavior of constraints will also be the subject of our modeling studies in Chapter 7.

4.2. Experiment 7: Effect of Verb Frame, Remnant, and Context on Gapping

We start our investigation of context effects on gradient grammaticality by providing experimental data on gapping, a phenomenon which has long been recognized as context-dependent in the theoretical literature, but which is under-researched from an experimental point of view. The present experiment will present some initial evidence for the fact that some constraints on gapping are context-dependent, i.e., that the effect of certain constraint violations disappears in an appropriate context.

4.2.1. Background

Gapping is a grammatical operation that deletes certain subconstituents of a coordinate structure. As examples consider (4.1)–(4.3) below, in which the (a) examples constitute gapped versions of the (b) examples:[1]

(4.1) a. I ate fish, Bill rice, and Harry roast beef.
b. I ate fish, Bill ate rice, and Harry ate roast beef. (Kuno 1976: (1))

(4.2) a. Tom has a pistol, and Dick a sword.
b. Tom has a pistol, and Dick has a sword. (Kuno 1976: (2))

(4.3) a. I want to try to begin to write a novel, and Mary $\left\{\begin{array}{r}\text{to try to begin to write}\\ \text{to begin to write}\\ \text{to write}\\ \emptyset\end{array}\right\}$ a play.
b. I want to try to begin to write a novel, and Mary wants to try to begin to write a play. (Kuno 1976: (3))

These examples indicate that gapping always deletes the matrix verb and leaves behind exactly two constituents as remnants (Kuno 1976: 318). Surveying previous work by Hankamer (1973),

[1] All examples in this section are taken from Kuno (1976).

Jackendoff (1971), and Ross (1970), Kuno (1976) also observes that certain functional principles affect the acceptability of gapping, such as the following restriction on the interpretation of the constituents left behind by gapping:[2]

(4.4) **The Minimal Distance Principle** [MINDIS]
The two constitutions left behind by Gapping can be most readily coupled with the constituents (of the same structures) in the first conjunct that were processed last of all. (Kuno 1976: (27))

The examples in (4.5) illustrate the Minimal Distance Principle: in (4.5a), the remnant *Tom* has to be paired with *Mary*, yielding the interpretation in (4.5b). It is not possible to pair *Tom* with the more distant subject *John*, yielding the interpretation in (4.5c).[3]

(4.5) a. John believes Mary to be guilty, and Tom to be innocent.
b. John believes Mary to be guilty, and John believes Tom to be innocent.
c. *John believes Mary to be guilty, and Tom believes Mary to be innocent. (Kuno 1976: (32))

A further generalization about gapping constructions is that the gap has to represent contextually given information, while the remnant has to constitute new information. Kuno (1976) captures this using the concept of Functional Sentence Perspective (FSP):

(4.6) **The FSP Principle of Gapping** [SENTP]
Constituents deleted by Gapping must be contextually known. On the other hand, the two constituents left behind by Gapping necessarily represent new information and, therefore, must be paired with constituents in the first conjunct that represent new information. [...] (Kuno 1976: (43))

Kuno (1976) notes that the FSP Principle seems to be able to override the Minimal Distance Principle. (4.7a) is acceptable as a gapped version of (4.7b), even though it violates MINDIS.

(4.7) a. With what did John and Billy hit Mary? John hit Mary with a stick, and Bill with a belt.
b. With what did John and Billy hit Mary? John hit Mary with a stick, and Bill hit Mary with a belt. (Kuno 1976: (34a))

Kuno's (1976) also observes that the remnants in a gapped sentences tend to be interpreted as a subject and its predicate:

[2] We supply constraint names for notational convenience.

[3] We use "?" or "*" to indicate the unacceptability of a given reading. On the meaning of acceptability marks in this thesis, see Section 3.1.6.

(4.8) **The Tendency for Subject-Predicate Interpretation** [SUBJPRED]
When Gapping leaves an NP and a VP behind, the two constituents are readily interpreted as constituting a sentential pattern, with the NP representing the subject of the VP. (Kuno 1976: (44))

This explains why (4.9a) can be interpreted as the gapped version of (4.9b) (where *Tom* is the subject of *donate*), but not as the gapped version of (4.9c) (where *Tom* is the subject of the object control verb *persuade*). Example (4.10a), on the other hand, not only has (4.10b) as a possible interpretation, but also (4.10c) (or at least (4.10c) is considerably better than (4.9c)). In (4.10c), *Tom* is the subject of *donate*, because the matrix verb *promise* is a subject control verb. Such a subject-predicate interpretation is preferred in gapping constructions. Note that (4.10c) violates MINDIS, thus indicating a competition between MINDIS and SUBJPRED.

(4.9) a. John persuaded Bill to donate $200, and Tom to donate $400.
b. John persuaded Bill to donate $200, and John persuaded Tom to donate $400.
c. *John persuaded Bill to donate $200, and Tom persuaded Bill to donate $400. (Kuno 1976: (47))

(4.10) a. John promised Bill to donate $200, and Tom to donate $400.
b. John promised Bill to donate $200, and John promised Tom to donate $400.
c. John promised Bill to donate $200, and Tom promised Bill to donate $400. (Kuno 1976: (48))

Finally, Kuno (1976) also observes that gapping cannot leave behind remnants that are part of a subordinate clause: (4.11a) cannot be understood as a gapped version of (4.11b).

(4.11) a. John persuaded Dr. Thomas to examine Jane and Bill Martha.
b. *John persuaded Dr. Thomas to examine Jane and Bill persuaded Dr. Thomas to examine Martha. (Kuno 1976: (52b))

This can be formulated as the generalization that the remnants in a gapping construction must be part of a simplex sentence:[4]

(4.13) **The Requirement for Simplex-Sentential Relationship** [SIMS]
The two constituents left over by Gapping are most readily interpretable as entering into a simplex-sentential relationship. The intelligibility of the gapped sentence declines drastically if there is no such relationship between the two constituents. (Kuno 1976: (54))

[4]It is not clear, however, how general this requirement is, see for instance the following example where gapping out of a PP complement seems to be possible:

(4.12) a. John gave Jane a picture of Elvis and Fred Bob Dylan.
b. John gave Jane a picture of Elvis and John gave Fred a picture of Bob Dylan.

Table 4.1: Main effects used to test the constraint set (Experiment 7)

verb frame (*Frame*)		remnant (*Remn*)	context (*Con*)
trans.	NP V NP	—	felicitous context
	NP V PP		null context (control)
	NP V VP		
	NP V PP-adj		
ditrans.	NP V NP NP	NP _ XP XP	felicitous context
	NP V NP PP	_ _ XP XP	null context (control)
	NP V NP VP	NP _ _ XP	
		NP _ XP _	

According to Kuno (1976: 316), "the Requirement for Simplex-Sentential Relationship is a very strong and nearly inviolable constraint", and a violation of this constraint leads to serious unacceptability. Kuno (1976) claims that the interaction of this constraint with weaker ones such as MINDIS, SENTP, and SUBJPRED, determines the degree of acceptability of gapped sentences. However, Kuno (1976) does not make this interaction explicit; he fails to give an account of how the degree of acceptability of a gapped sentence is computed from the constraint violations it incurs. The present experiment aims to overcome this limitation. Using experimental data we investigate how the interaction of constraints on gapping determines the degree of acceptability of a gapped structure.

Gapping is an under-researched area in psycholinguistics; a small number of judgment experiments on ellipsis in coordinated structures were reported by Greenbaum (1977), Greenbaum and Meyer (1982), and Meyer (1979). More recently, Carlson (1999) presented an experimental investigation of the effect of parallelism and prosody on the preferred interpretation of a gapped sentence. None of these studies dealt with context effects, the focus of Experiments 7 and 8.

4.2.2. Introduction

Experiment 7 was designed to investigate whether certain constraints on gapping that have been proposed in the literature have a gradient effect on the acceptability of gapped sentences: (a) the verb frame of the gapped verb (b) whether the remnant left behind by gapping is a complement or an adjunct, (c) the structure of the remnant, and (d) the context preceding the gapped sentence. Table 4.1 gives an overview of the factors included in this experiment and their levels.

The factor verb frame (*Frame*) included both transitive and ditransitive verbs. The transitive case included verbs with NP, PP, and VP complements. PP adjuncts were also included in order to test the claim that adjunct remnants are more acceptable than complement remnants (Hankamer 1973). The following examples illustrate the levels of the factor *Frame* for

transitive verbs:

(4.14) a. **NP V NP:** She repeated the question, and he the answer.
b. **NP V PP:** She negotiated with the manager, and he with the secretary.
c. **NP V VP:** She expected to win, and he to lose.
d. **NP V PP-adj:** She read in the bedroom, and he in the lounge.

For ditransitive verbs, the factor *Frame* included verbs that have an NP as their first complement, and another NP, a PP, or a VP as their second complement, such in (4.15).

(4.15) a. **NP V NP NP:** She charged the client 50 pounds, and he the manufacturer 100 pounds.
b. **NP V NP PP:** She accompanied the boy to school, and he the girl to university.
c. **NP V NP VP:** She authorized the manager to leave, and he the secretary to stay.

Transitive verbs allow only one type of remnant (where the subject and the object are left behind, while the verb is gapped). Ditransitive verbs, on the other hand, allow more complicated remnants, which we took into account by including the additional factor remnant type (*Remn*) for ditransitive verbs. The levels of *Remn* can be exemplified by the following sentences:

(4.16) a. **NP _ XP XP:** She charged the client 50 pounds, and he the manufacturer 100 pounds.
b. **_ _ XP XP:** She charged the client 50 pounds, and the manufacturer 100 pounds.
c. **NP _ _ XP:** She charged the client 50 pounds, and he 100 pounds.
d. **NP _ XP _:** She charged the client 50 pounds, and he the manufacturer.

Note that we use pronouns in (4.16c) and (4.16d) to make sure that the remnant is interpreted as the subject NP.

Context (*Con*), the third factor in the experiment, was meant to test the influence of context on the acceptability of gapping. A felicitous context for gapping (according to Kuno's 1976 SENTP constraint) is one in which the gapped constituent contains given information, while the remnants constitute new information. Such a given-new partition can be realized using a question context: new constituents in the answer are realized as *wh*-phrases in the question, while given constituents in the answer are realized as full NPs in the question. This is illustrated by the questions in (B.38)con, which constitute felicitous contexts for the transitive sentences in (4.14):

(4.17) a. What did Hanna and Michael repeat?
b. Who did Emily and Matthew negotiate with?
c. What did Rachel and Andrew expect to do?
d. Where did Rebecca and Mark read?

The factor *Con* was the same for the ditransitive condition. Here are the felicitous contexts for the examples in (4.16):

(4.18) a. Who did Hanna and Michael charge what?
b. Who did Hanna charge what?
c. What did Hanna and Michael charge the client?
d. Who did Hanna and Michael charge 50 pounds?

A null context condition was included as a control condition, allowing us to determine how subjects behave in the absence of contextual information. (The influence of non-felicitous and neutral contexts on gapping was investigated in Experiment 8.)

4.2.3. Predictions

As far as the factor *Frame* is concerned, no clear predictions can be derived from the literature as to the effect of complement type (NP, PP, or VP) or arity (transitive or ditransitive) of the verb. As for the complement/adjunct status of the remnant, the experiment allows us to test Hankamer's (1973) claim that PPs adjuncts are more acceptable than PP complements.[5]

For the factor *Remn*, the constraint MINDIS predicts that the remnant _ _ XP XP is more acceptable than the remnants NP _ _ XP and NP _ XP _. Another relevant prediction is that the remnant NP _ XP XP is unacceptable, based on Kuno's (1976: 318) claim that gapping has to leave behind exactly two constituents.

As for the effect of *Con*, Kuno's (1976) constraint SENTP predicts that the acceptability of a gapped sentence should be increased in a felicitous context, compared to the control condition (the null context).

Furthermore, we predict an interaction between the factors *Remn* and *Con*, based on Kuno's (1976) observation that the satisfaction of SENTP seems to override a violation of the MINDIS (see Section 4.2.1).

4.2.4. Method

4.2.4.1. Subjects

Fifty-five native Speakers of English from the same population as in Experiment 4 participated in the experiment.

The data of two subjects were excluded because they were bilingual (by self-assessment). The data of a further two subjects were excluded because they were linguists

[5] Consider the following examples from Hankamer (1973), which are analogous to our sentences (4.14b) and (4.14d) (the acceptability judgments are his):

(4.19) a. *Max wanted to put the eggplant on the table, and Harvey in the sink.
b. ?Max writes plays in the bedroom, and Harvey in the basement.

(by self-assessment). The data of another two subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 49 subjects for analysis. Of these, 29 subjects were male, 20 female; eight subjects were left-handed, 41 right-handed. The age of the subjects ranged from 14 to 52 years, the mean was 30.6 years.

4.2.4.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials The experiment included two subdesigns, as illustrated in Table 4.1. For the transitive items, a full factorial design was used with verb frame (*Frame*) and context (*Con*) as the two factors. (4.14) gives example stimuli for the transitive condition; (4.15) givens examples for the ditransitive condition. Example contexts are given in (4.17). This yielded a total of $Frame \times Con = 4 \times 2 = 8$ cells. For the ditransitive items, the additional factor remnant type (*Remn*) was included, yielding $Frame \times Remn \times Con = 3 \times 4 \times 2 = 24$ cells Four lexicalizations were used for each of the cells, which resulted in a total of 128 stimuli.

A set of 32 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

To control for possible effects from lexical frequency, the stimuli in both subdesigns were matched for frequency. Verb and noun frequencies were obtained from a lemmatized version of British National Corpus (100 million words) and average frequencies were computed for the verb, the head noun of the subject, and the head noun of the object for each frame. An ANOVA confirmed that the average verb, subject, and object frequencies did not differ significantly between frames.

4.2.4.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions We used an English version of the instructions in Experiment 1. Where contextualized stimuli were presented, subjects were told that each sentence would be presented in context, defined as a single sentence preceding the target sentence. Subjects were instructed to judge the acceptability of the target sentence, and to take the context into account in their judgments. The task was illustrated by examples.

Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

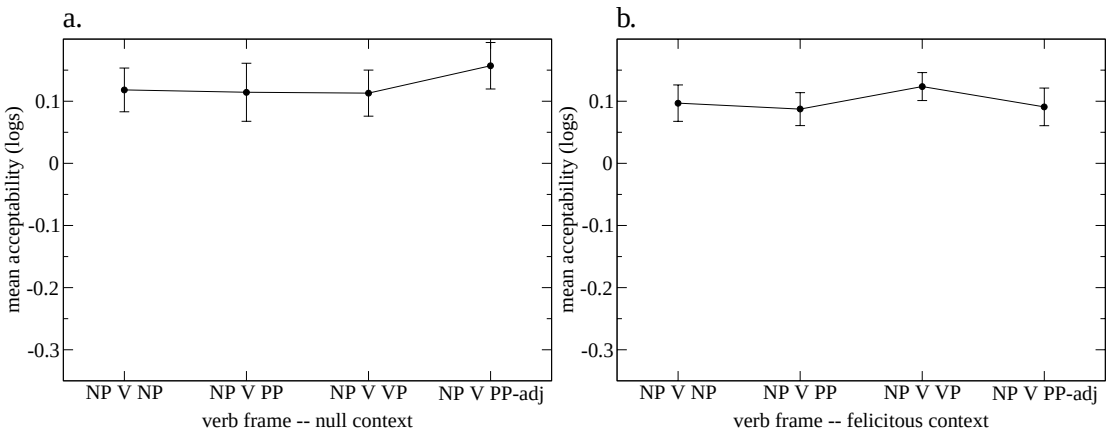

Figure 4.1: Mean judgments for gapping by verb frame and context, transitive frames (Experiment 7)

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

A between subjects design was used to administer the factor *Con*: subjects in Group A judged non-contextualized stimuli, while subjects in Group B judged contextualized stimuli. The factors *Frame* and *Remn* were administered within subjects.

For each group, four test sets were generated: each test set contained one lexicalization for each of the 16 cells in the design. Lexicalizations were assigned to test sets using Latin squares. Four separate Latin squares were applied: two for the transitive condition (null context and context) and two for the ditransitive condition (null context and context).

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 32 test items: 16 experimental items and 16 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to a group and a test set; 26 subjects were assigned to group A, and 23 to group B. Instructions, examples, training items, and fillers were adapted for Group B to take context into account.

4.2.5. Results

The data were normalized as in Experiment 1. Separate analyses of variance (ANOVAs) were performed for the transitive and ditransitive verb frames. The analysis for the transitive frames failed to find a significant main effect of verb frame. The main effect of context was significant only by items ($F_1(1,47) = .326$, $p = .571$; $F_2(1,6) = 29.720$, $p = .002$), and the interaction of frame and context was non-significant. The average judgments for the transitive condition are graphed in Figure 4.1.

For the ditransitive frames, a marginal main effect of verb frame was found

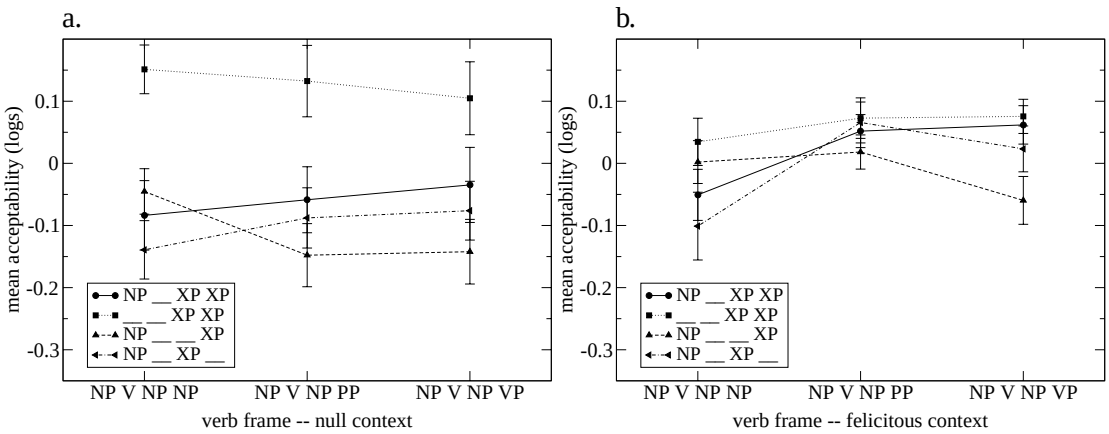

Figure 4.2: Mean judgments for gapping by verb frame, remnant type, and context, ditransitive frames (Experiment 7)

($F_1(2,94) = 2.727$, $p = .071$; $F_2(2,12) = 6.037$, $p = .015$). Furthermore, the ANOVA showed a highly significant main effect of remnant type ($F_1(3,141) = 18.936$, $p < .0005$; $F_2(3,18) = 6.564$, $p = .003$), and an interaction of verb frame and context ($F_1(2,94) = 5.661$, $p = .005$; $F_2(2,12) = 5.096$, $p = .025$). The interaction of remnant type and context was significant by subjects ($F_1(3,141) = 5.483$, $p = .001$; $F_2(3,18) = 1.847$, $p = .175$). Finally, there was an interaction of remnant type and verb frame, significant by subjects and marginal by items ($F_1(6,282) = 3.817$, $p = .001$; $F_2(6,36) = 1.972$, $p = .096$). No main effect of context was found, and all the remaining interactions were non-significant.

The mean judgments for the null context conditions are graphed in Figure 4.2a. This graph shows that the _ _ XP XP remnant is more acceptable than the other remnants. This effect is consistent across all frame types. A comparison with Figure 4.2b (showing the mean judgments for the context condition) demonstrates that the remnant effect disappears in a felicitous context: the _ _ XP XP remnant is not more acceptable than the other remnants.

To verify this observation, we carried out a post-hoc Tukey test on the *Remn*/*Con* interaction. In the null context condition, we found that the _ _ XP XP is significantly more acceptable than all other remnants (by subjects only, $\alpha < .05$ in all three cases). None of the other remnants were significantly different from each other. In the context condition, on the other hand, we found that all remnants were equally acceptable.

Another post-hoc Tukey test was carried out to investigate the *Frame*/*Con* interaction. In the null context, none of the frames were significantly different from each other. In the context condition, we found that the NP V NP NP frame was significantly less acceptable than the NP V NP PP frame (by subjects only, $\alpha < .05$). Furthermore, we found that the NP V NP PP frame was significantly more acceptable in the context condition than in the null context condition (by subjects only, $\alpha < .05$). The same effect was found for the NP V NP VP frame (by subjects only, $\alpha < .05$).

We did not investigate the *Frame*/*Remn* interaction any further, as it is simply a by-product of the *Remn*/*Con* and the *Frame*/*Con* interactions, already discussed above.

4.2.6. Discussion

For transitive verbs, we found that gapping is equally acceptable for all types of verbal complements tested (NP, PP, VP). We also failed to find a difference between PP complements and PP adjuncts. This result settles the controversy on the status of complements and adjuncts in gapping: Hankamer (1973) claims that PP adjuncts are more acceptable than PP complements, a claim that is disputed by Jackendoff (1971) and Kuno (1976). These negative results are also important the follow-up experiment on gapping (Experiment 8), as they make it possible to disregard the distinction between different verb frames, and between adjuncts and complements, thus enabling us to use a more compact experimental design.

In contrast to transitive verbs, ditransitive verbs showed an effect of *Frame*: in a felicitous context, the NP V NP NP frame was less acceptable than the NP V NP PP frame. Also, the acceptability of the NP V NP PP and NP V NP VP frames was found to be context-dependent. Note however that these effects, for which the literature on gapping fails to offer an explanation, are rather small (see Figure 4.2).

The main finding of Experiment 7 is the effect of remnant type and its interaction with context. We showed that the _ _ XP XP remnant is more acceptable than all the other remnants, an effect that is very strong in a null context, but disappears completely in a felicitous context. This provides convincing evidence for Kuno's (1976) Minimal Distance Principle, and in particular for his observation that a violation of MINDIS can be overridden by satisfying the context requirements on gapping (his constraint SENTP).

On the other hand, we found that the NP _ XP XP remnant is not significantly less acceptable than NP _ _ XP and NP _ XP _, contrary to Kuno's (1976) claim that gapping must leave behind exactly two remnants. The finding is consistent with observations by Steedman (1990), who argues against Kuno's two-remnant restriction.

Now let us briefly consider an alternative explanation for the interaction of remnant type and context. One could argue that this effect is actually due to the contexts used, rather than to the stimulus sentences proper. Some initial plausibility for this view derives from the fact that two of the remnants (NP _ XP XP and _ _ XP XP) used double *wh*-question as contexts (see (4.18a) and (4.18b)), while the other two remnants (NP _ _ XP and NP _ XP _) had single *wh*-question as contexts (see (4.18c) and (4.18d)). It seems plausible to assume that multiple *wh*-questions are less acceptable than single ones, and maybe subjects took the acceptability of the context into account when they judged the acceptability of the stimulus sentences.

To test this hypothesis, an ANOVA was conducted on the contextualized data with question type as the only factor. This yielded an effect of question type which was significant by

subjects ($F_1(1,2) = 8.982$, $p = .007$; $F_2(1,3) = 1.257$, $p = .344$). However, this effect went the other way than was expected: single questions (mean $= -.0085$) were less acceptable than double questions (mean $= .0410$). This result allows us to rule out the hypothesis that the effect of *Remn* is due to the type of question used, rather than to the remnant itself.

Another alternative explanation for the remnant is that _ _ XP XP is more acceptable because it does not contain a subject pronoun. This pronoun is present in the other three remnants and might reduce acceptability in the null context condition, as it cannot be anchored to an NP in the context. This would explain why the remnant effect disappears the in context, where such an antecedent is provided (see (4.14) and (4.17)). This alternative explanation for the remnant effect cannot be ruled out on the basis of Experiment 7. We will address this issue in the next experiment, where we investigate the behavior of gapping in non-felicitous contexts. A non-felicitous context provides an antecedent for the subject pronoun, but differs from a felicitous context in that it violates SENTP.

4.2.7. Conclusions

The results of Experiment 7 confirms the usefulness of an experimental approach to linguistic data by applying magnitude estimation to gapping constructions. Experiment 7 showed that PP adjuncts and PP complements are equally acceptable as remnants in gapping, a fact that was surrounded by controversy in the theoretical literature. It also provided evidence against the claim that gapping must leave behind exactly two remnants (Kuno 1976). Another theoretically interesting result is that subject remnants are less acceptable than object remnants, an effect that turned out to be context-dependent.

Context effects such as this one are the focus of the remainder of this chapter. In the next experiment, we will extend our investigation of context effects in gapping and arrive at the hypothesis that soft constraints are context-dependent, while hard constraints are context-independent, i.e., immune to context effects (see Section 4.1.1). In Experiments 9–12 we will then ask if this hypothesis is borne out with respect to two phenomena already investigated in Chapter 3, viz., extraction and word order.

4.3. Experiment 8: Effect of Remnant, Subject-Predicate, Simplex S, and Context on Gapping

Based on the findings of Experiment 7, the present experiment will provide a more systematic investigation of context effects on gapping. We will report experimental data on the interaction of three different constraints on gapping and determine the ranking of these three constraints. Moreover, we investigate which of these constraints are subject to context effects, and which ones are context-independent.

Table 4.2: Main effects used to test the constraint set (Experiment 8)

MINDIS (*Dis*)	SUBJPRED (*Pred*)	SIMS (*Sim*)	SENTP (*Con*)
not violated (_ _ XP XP)	not violated	not violated	not violated (fel. context)
violated (NP _ _ XP)	violated	violated	violated (non-fel. context)
			neutral context (control)
			null context (control)

4.3.1. Introduction

Table 4.2 gives an overview of the factors included in Experiment 8. The constraints we investigate are the ones detailed in Section 4.2.1, either violated or not: Minimal Distance (MINDIS), Functional Sentence Perspective (SENTP), Subject-Predicate Interpretation (SUBJPRED), and Simplex-Sentential Relationship (SIMS).

The constraint MINDIS (see (4.4)) is satisfied if the distance between the remnants and their antecedents is minimal, as in (4.20a), where *the thief* can be paired with *the criminal* and *for robbing the bank* can be paired with *for burgling the house*. (4.20b), on the other hand, is in violation of MINDIS, as *she* cannot be paired with *the neighbor*, but has to be paired with the subject *he*.

(4.20) a. He punished the criminal for robbing the bank and the thief for burgling the house.
b. He helped the neighbor by doing the shopping and she by washing the dishes.
c. He punished the criminal for robbing the bank and the thief the house.
d. He helped the neighbor by doing the shopping and the friend by washing the dishes.

Another constraint on gapping postulated by Kuno (1976) is SUBJPRED (see (4.8)), which requires that the remnants left behind by gapping are interpreted as a subject and its predicate. This constraint is met in (4.20a), where *the thief* is the subject of *for burgling the house*, but it is violated in (4.20d), where the subject of *washing the dishes* is not the remnant *the friend*, but the main clause subject *he*.

The constraint SIMS (see (4.13)) requires that the constituents left behind by gapping have to be part of a simplex sentence, i.e., gapping out of subordinate clauses is disallowed. This constraint is met in (4.20a), where the gapped clause is interpreted as *he punished the thief for burgling the house*, while it is violated in (4.20c), where the interpretation of the gapped clause is *he punished the thief for robbing the house*.

Finally, the experiment included the constraint SENTP (see (4.6)), which governs the context required for gapping. Extending the results of Experiment 7, we included not only a felicitous context condition, where the remnants are new, while the gap is given (i.e., SENTP is satisfied), but also a non-felicitous context, where the remnants are given, while the gap is new (i.e., SENTP is violated). The contexts were formulated as questions, on par with Experiment 7.

In addition to the felicitous and non-felicitous contexts, we included two control conditions: a null context condition and a neutral context condition. In the null context condition, the stimuli were presented in isolation. In the neutral context condition, the stimuli were prefixed by the question *What happened?*, which indicates an all focus Information Structure.

The examples in (4.21) show the felicitous contexts that belong to the stimuli in (4.20), while (4.22) gives the corresponding non-felicitous contexts.

(4.21) a. Who did Michael punish, and why?
b. How did David and Hanna help the neighbor?
c. Who did Michael punish, and why?
d. Who did David help, and how?

(4.22) a. Why did Michael punish the criminal and the thief?
b. Who did David and Hanna help, and how?
c. Why did Michael punish the criminal and the thief?
d. How did David help the neighbor and the friend?

4.3.2. Predictions

4.3.2.1. Constraints

The results of Experiment 7 and the claims in the theoretical literature on gapping provide a set of predictions regarding the constraints investigated in the present experiment.

We expect strong unacceptability for a violation of SIMS, i.e., for sentences where the remnants are not in a simplex-sentential relationship. Intuitively, a violation of SIMS is so serious that it cannot be remedied by the satisfaction of other constraints such as MINDIS, SUBJPRED, or SENTP.

An effect of MINDIS is also predicted, i.e., structures with subjects remnants (see (4.20b)) are expected to be reduced in acceptability. In line with the findings of Experiment 7 this effect should disappear in a felicitous context (see (4.21b)).

We also expect a significant effect of SUBJPRED; gapped sentences that do not allow a subject-predicate interpretation of the remnants (see (4.20d)) are predicted to be dispreferred. In line with Kuno's (1976) observations, we expect this effect to interact with MINDIS, and possibly with SENTP, i.e., with context (even though Kuno (1976) does not explicitly mention this possibility).

Finally, Kuno's (1976) account also predicts an effect of SENTP, i.e., a felicitous context should improve the overall acceptability of a gapped sentence.

4.3.2.2. Constraint Ranking

Chapter 3 provided evidence for a classification of constraints into soft and hard constraints. Soft constraints cause gradient acceptability effects, while hard constraints induce binary ac-

ceptability judgments.

This leads to the prediction that the constraints tested in this experiment cluster into hard and soft constraints. Hard constraints are expected to receive a high ranking, i.e., trigger a high degree of unacceptability; while soft constraints will receive a low ranking, i.e., cause only mild unacceptability when violated.

Intuitively, SIMS is a good candidate for a hard constraint, while SUBJPRED and MINDIS are probably soft constraints. A particularly interesting question is how context interacts with soft and hard constraints. It seems plausible to expect soft constraints to be more susceptible to context effects than hard ones.

These predictions will be tested using a series of planned comparisons to determine if the constraint violations differ in the relative degree of ungrammaticality that they cause.

4.3.2.3. Constraint Interaction

Another important finding in Chapter 3 was that constraint violations are cumulative, i.e., that the degree of unacceptability of a sentence increases with the number of constraint violations it incurs. We expect the cumulativity of violations to be in evidence in the present experiment. Again a set of planned comparisons will be used to test this prediction.

4.3.3. Method

4.3.3.1. Subjects

Sixty native speakers of English from the same population as in Experiment 4 participated in the experiment. None of them had previously participated in Experiment 7.

The data of two subjects were excluded because they were linguists (by self-assessment). The data of another three subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 55 subjects for analysis. Of these, 32 subjects were male, 23 female; eight subjects were left-handed, 47 right-handed. The age of the subjects ranged from 17 to 72 years, the mean was 31.6 years.

4.3.3.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials A full factorial design was used which included the factors *Dis*, *Sim*, *Pred*, and *Con*, representing the constraints MINDIS, SIMS, SUBJPRED, and SENTP, respectively (see Table 4.2 for an overview of the experimental design). The factors *Dis*, *Sim*, and *Pred* had two levels (constraint violated or not violated), while the factor *Con* had four levels: constraint

violated (non-felicitous context), not violated (felicitous context), plus the two control conditions (null context and neutral context). (4.20) lists example stimuli; example contexts are given in (4.21) and (4.22). This yielded a total of $Dis \times Sim \times Pred \times Con = 2 \times 2 \times 2 \times 4 = 32$ cells. Eight lexicalizations were used for each of the cells, which resulted in a total of 256 stimuli.

A set of 24 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

The materials were matched for frequency using the same procedure as in Experiment 7.

4.3.3.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions We used an English version of the instructions in Experiment 1. Where contextualized stimuli were presented, subjects were told that each sentence would be presented in context, defined as a single sentence preceding the target sentence. Subjects were instructed to judge the acceptability of the target sentence, and to take the context into account in their judgments. The task was illustrated by examples.

Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

A between subjects design was used to administer the factor *Con*: subjects in Group A judged non-contextualized stimuli, while subjects in Group B judged contextualized stimuli.

For Group A, four test sets were used: each set contained two lexicalizations for each of the cells in the design $Dis \times Sim \times Pred$, i.e., a total of 16 items. For Group B, eight test sets were used, each set containing one lexicalization and three contextualizations for each cell, i.e., a total of 24 items. Lexicalizations were assigned to test sets using Latin squares. Two separate Latin squares were applied: one for the null context condition and one for the context condition.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. In Group A, each subject saw 32 items: 16 experimental items and 16 fillers. In Group B, each subject saw 40 items: 24 experimental items and 16 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to a group and a test set; 25 subjects were assigned to Group A, 30 to Group B. Instructions, examples, training items, and fillers were adapted for Group B to take context into account.

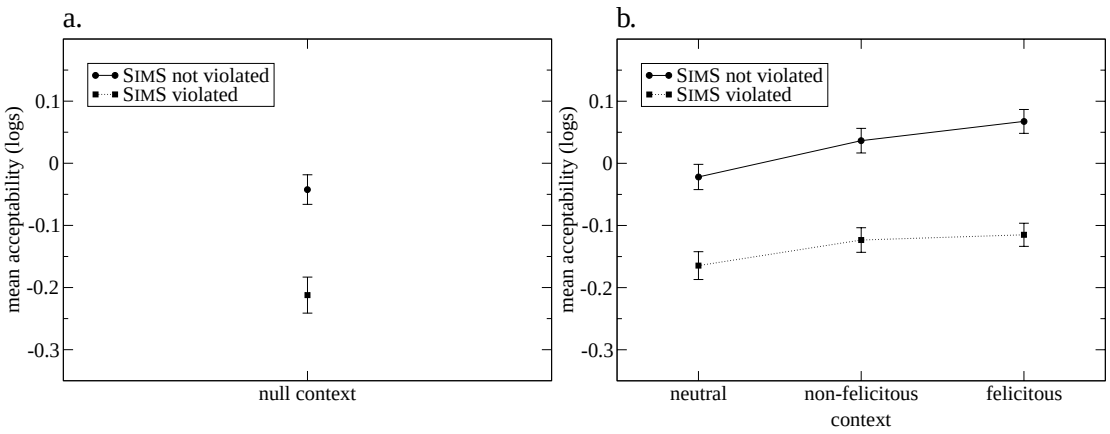

Figure 4.3: Context effects for SIMS (Experiment 8)

4.3.4. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

4.3.4.1. Constraints

Simplex Sentence In the null context condition, a highly significant main effect of *Sim* was found ($F_1(1,24) = 23.415$, $p < .0005$; $F_2(1,7) = 18.918$, $p = .003$). The same effect of *Sim* was present in the context condition ($F_1(1,29) = 97.310$, $p < .0005$; $F_2(1,7) = 15.548$, $p = .006$). The interaction between *Sim* and context was non-significant.

Figure 4.3 depicts the mean judgments for a violation of SIMS in all contexts. It indicates that SIMS violations have a strong effect on acceptability and illustrates the absence of a context effect: a violation of SIMS results in the same decrease of acceptability in all contexts (including the null context and the neutral context).

Minimal Distance In the null context condition, a highly significant main effect of *Dis* was found ($F_1(1,24) = 25.997$, $p < .0005$; $F_2(1,7) = 14.612$, $p = .007$). *Dis* was also significant in the context condition ($F_1(1,29) = 23.315$, $p < .0005$; $F_2(1,7) = 11.421$, $p = .012$), where an interaction of *Dis* and *Sim* was also present, significant by subjects only ($F_1(1,29) = 4.568$, $p = .001$; $F_2(1,7) = 2.111$, $p = .190$).

The ANOVA also revealed a significant interaction of *Dis* and context ($F_1(2,58) = 4.568$, $p = .014$; $F_2(2,14) = 6.553$, $p = .010$), which is depicted in Figure 4.4. A post-hoc Tukey test was carried out to determine the locus of this interaction. The effect of *Dis* was significant in the neutral context and in the non-felicitous context (by subjects only, $\alpha < .05$ in both cases). However, no significant effect of *Dis* was found in the felicitous context. This demonstrates that the effect of *Dis* disappears in the felicitous context, in line with our predic-

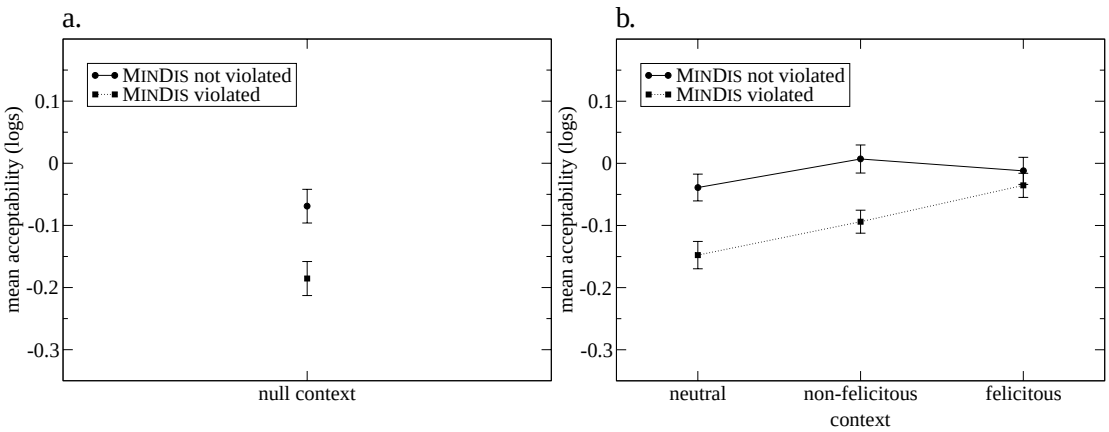

Figure 4.4: Context effects for MINDIS (Experiment 8)

tions.

Subject-Predicate Interpretation The main effect of *Pred* failed to reach significance in the null context condition. In the context condition, a main effect of *Pred* was found ($F_1(1,29) = 19.377$, $p < .0005$; $F_2(1,7) = 9.891$, $p = .016$). The interaction of *Pred* and context failed to be significant. There was, however, an interaction of *Pred* and *Sim* that was significant by subjects only ($F_1(1,29) = 11.453$, $p = .002$; $F_2(1,7) = 2.524$, $p = .156$). Figure 4.5 depicts the effect of *Pred* for each context.

The presence of a *Pred*/*Sim* interaction might indicate that the effect of *Sim* blocks the context effect of *Pred*. Recall that a violation of SIMS leads to a high degree of unacceptability, while SUBJPRED only has a small effect on acceptability. It is therefore appropriate to factor out violations of SIMS (and other constraints), and to look at the effect of context on single violations of SUBJPRED. The mean judgments for single violations of SUBJPRED are depicted in Figure 4.6, which indicates that the effect of *Pred* in the neutral context is stronger than in the other contexts.

To confirm this observation, we conducted a series of planned comparisons on the single violations of SUBJPRED for the four context conditions. The significance level was adjusted using the Bonferroni procedure, i.e., we set $p = .0125$. In the null context, the felicitous context, and the non-felicitous context, no significant effect of a single SUBJPRED violations was found. In the neutral context, however, a single violation of SUBJPRED lead to a significant reduction in acceptability (by subjects only, $F_1(1,29) = 8.327$, $p = .007$; $F_2(1,7) = 5.610$, $p = .050$).

Functional Sentence Perspective The ANOVA on the context condition showed a significant main effect *Con* ($F_1(1,29) = 10.209$, $p < .0005$; $F_2(1,7) = 13.082$, $p = .001$). A post-hoc Tukey test was conducted to investigate the locus of the *Con* effect. It was found that the neutral context was significantly less acceptable than both the felicitous and the non-felicitous

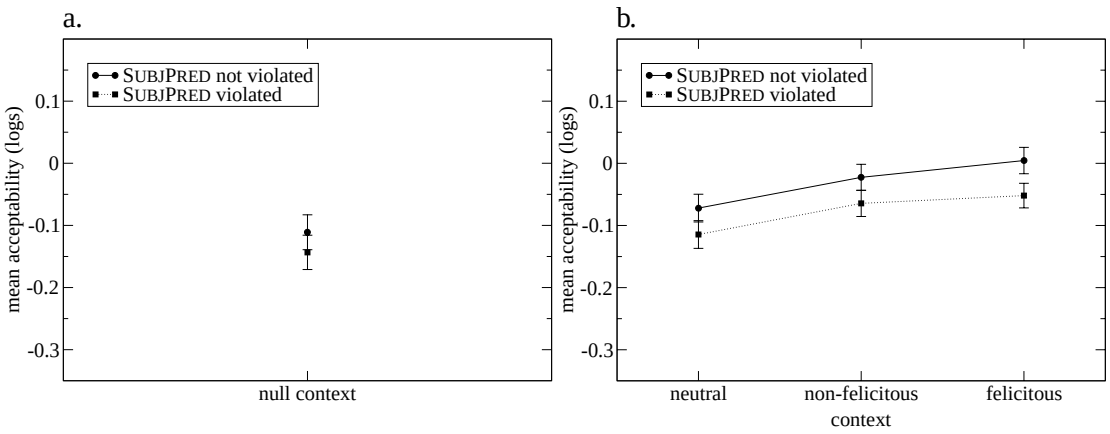

Figure 4.5: Context effects for SUBJPRED (Experiment 8)

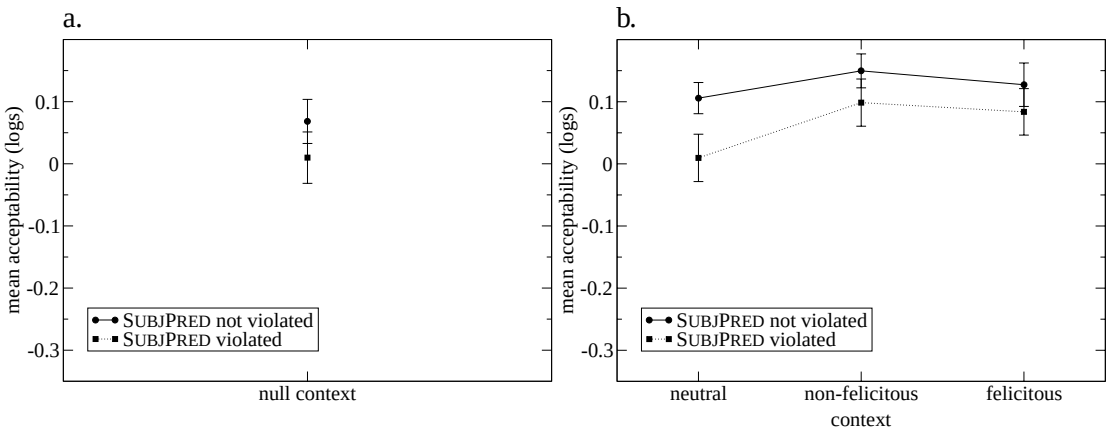

Figure 4.6: Context effects for SUBJPRED, single violations (Experiment 8)

context ($\alpha < .01$ in both cases). However, there was no difference between the felicitous and the non-felicitous context.

4.3.4.2. Constraint Ranking

To establish constraint ranking, we carried out further tests on single violations of SIMS, MINDIS, and SUBJPRED. Due to the context effects for MINDIS and SUBJPRED reported above, such tests are not meaningful for all contexts. Recall that we found that a violation of MINDIS disappears in the felicitous context; a violation of SUBJPRED failed to be significant in the felicitous and the non-felicitous context. This means that an analysis by constraint type should only be conducted in the neutral context and in the null context.

Figure 4.7 compares the degree of unacceptability caused by single violations of the constraints SIMS, MINDIS, and SUBJPRED. The graph indicates that a violation of SUBJPRED only has a small effect on acceptability. A violation of SIMS leads to serious unacceptability,

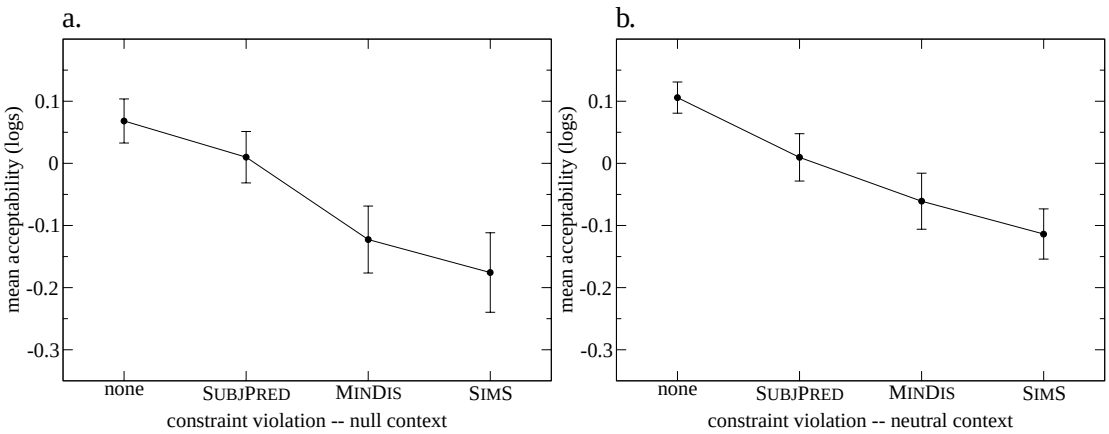

Figure 4.7: Constraint ranking, single violations (Experiment 8)

while a violation of MINDIS is somewhere in between.

To test if these differences in unacceptability were significant, we conducted a series of planned comparisons on the subset of the data that only contained single violations. Three planned comparisons were carried out for the null context, and three for the neutral context. Again, a Bonferroni adjustment was uses, i.e., the significance level was set at $p = .0167$.

In the null context, we found a significant difference between a single violation of SUBJPRED (mean $= .0099$) and a single violation of MINDIS (mean $= -.1226$) (by subjects only, $F_1(1,24) = 8.533$, $p = .007$; $F_2(1,7) = 6.113$, $p = .043$). Also the difference between a SUBJPRED violation and a SIMS violation (mean $= -.1757$) was significant (by subjects only, $F_1(1,24) = 10.338$, $p = .004$; $F_2(1,7) = 5.594$, $p = .050$). There was no significant difference, however, between a MINDIS violation and a SIMS violation.

In the neutral context we failed to find a significant difference between a single violation of SUBJPRED (mean $= .0096$) and a single violation of MINDIS (mean $= -.0609$). The difference between a SUBJPRED violation and a SIMS (mean $= -.1137$) violation was significant (by subjects only, $F_1(1,29) = 11.824$, $p = .002$; $F_2(1,7) = 2.814$, $p = .137$). There was no significant difference between a MINDIS violation and a SIMS violation.

The results of these planned comparisons are compatible with overall constraint ranking of $\{\text{SIMS}, \text{MINDIS}\} \gg \text{SUBJPRED}$ (recall that "$\gg$" means "is ranked higher than").

4.3.4.3. Constraint Interaction

To test the hypothesis that constraint violations are cumulative, we carried out a set of planned comparisons on the null context and the neutral context (recall that all three constraint violations were observed only in these contexts, allowing for a full evaluation of cumulativity effects). As in Experiments 4–6, we computed the mean acceptability for stimuli incurring zero violations (one sentence type), one violation (three sentence types), two violations (three

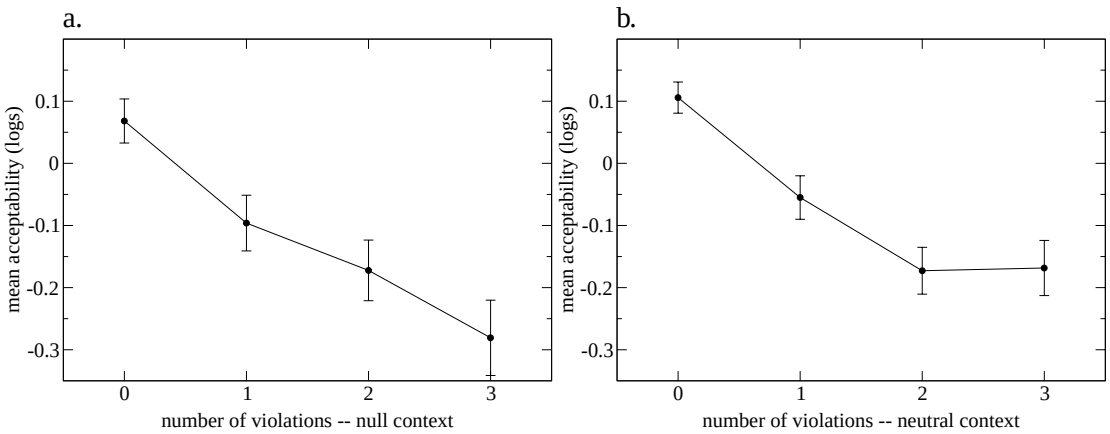

Figure 4.8: Cumulativity of constraint violations (Experiment 8)

sentence types), and three violations (one sentence types). The resulting means are graphed in Figure 4.8. Three planned comparisons were carried on each data set, hence a Bonferroni adjustment yielded $p = .0167$ as the significance level.

In the null context a planned comparison of zero violations (mean $= .0682$) and a single violation (mean $= -.0961$) yielded a significant effect ($F_1(1,24) = 13.511$, $p = .001$; $F_2(1,7) = 14.273$, $p = .007$). We also discovered a difference between a single and a double violation (mean $= -.1723$) (significant by subjects and marginal by items, $F_1(1,24) = 17.982$, $p < .0005$; $F_2(1,7) = 9.364$, $p = .018$), and between a double and a triple violation (mean $= -.2808$) (significant by subjects and marginal by items, $F_1(1,24) = 10.128$, $p = .004$; $F_2(1,7) = 5.887$, $p = .046$).

This pattern was replicated in the neutral context condition, where we found a significant difference between a zero violations (mean $= .1058$) and a single violation (mean $= -.0550$) ($F_1(1,29) = 29.779$, $p < .0005$; $F_2(1,7) = 26.240$, $p = .001$). The difference between a single and a double (mean $= -.1728$) violation was significant by subjects and marginal by items ($F_1(1,29) = 11.059$, $p = .002$; $F_2(1,7) = 8.657$, $p = .022$). However, there was no significant difference between a double violation and a triple violation (mean $= -.1684$).

4.3.5. Discussion

4.3.5.1. Constraints

Experiment 8 found main effects of *Dis*, *Pred*, *Sim*. This demonstrated that violations of the constraints MINDIS, SUBJPRED, and SIMS significantly reduce the acceptability of gapped sentences, as predicted by Kuno's (1976) account of gapping. A main effect of *Con* was also present, but contrary to predictions, no difference between the acceptability of gapping in a felicitous and a non-felicitous context was found. However, the acceptability of gapping in the felicitous and the non-felicitous context was significantly higher than in the neutral context.

This seems to indicate that even a non-felicitous context provides an Information Structure that is partially compatible with the requirements of the constraint SENTP.

We also found that SENTP interacts with other constraints on gapping. A significant interaction of *Con* and *Dis* was obtained: a violation of MINDIS leads to reduced acceptability in the null context, the neutral context, and the non-felicitous context. In the felicitous context (that satisfies the information structure constraint SENTP), the effect of *Dis* disappeared. Note that the null context and the neutral context behaved in the same fashion with respect to MINDIS violations; this is expected based on the hypothesis that even a null context carries implicit information structural assumptions, and is interpreted by subjects on par with a neutral (all focus) context.

Like the *Dis* effect, the effect of *Pred* was also found to be context-dependent. Considering stimuli that incur a single violation of SUBJPRED, we found a significant effect of *Pred* only in the neutral context; in the felicitous and non-felicitous context, the effect of *Pred* was too small to be significant. Also, in the null context, no effect of *Pred* was found, even though this would be expected under the assumption that the null context behaves like an neutral (all focus) context.

In contrast to MINDIS and SUBJPRED, the Simplex S constraint (SIMS) was found to be immune to context effects: it caused consistently strong unacceptability, independent of which context was presented. This is in line with our predictions regarding the behavior of SIMS.

Another one of Kuno's (1976) observations can be tested against the data from Experiment 8. Examples like (4.9) and (4.10) seem to indicate that a satisfaction of SUBJPRED can override a violation of MINDIS. However, we failed to find an interaction of *Dis* and *Pred* in either the null context condition or the context condition. This might indicate that the interaction of SUBJPRED and MINDIS that Kuno (1976) observes is limited to examples like the ones in (4.9) and (4.10), and does not generalize to our experimental stimuli.

Finally, the results of the present experiment allow us to evaluate the alternative explanation for the *Dis* effect we discussed in Section 4.2.6: the _ _ XP XP remnant is more acceptable than the XP _ _ XP remnant because the latter contains a subject pronoun, which reduces acceptability if it is not contextually anchored (in a null or neutral context). This explanation can be ruled out on the basis of Experiment 8, which demonstrated a *Dis* effect for the non-felicitous context condition, i.e., even if the subject pronoun can be anchored to a contextually given NP.

4.3.5.2. Constraint Ranking

A second set of predictions for Experiment 8 was based on Chapter 3, where we arrived at the hypothesis that there are two types of constraints: hard constraint that lead to serious unacceptability and soft constraints that cause only mild unacceptability. Consider Figure 4.7,

which graphs the unacceptability incurred by single violations of the three constraints SIMS, MINDIS, and SUBJPRED. We found that a SIMS and MINDIS violations were significantly more serious than a violation of SUBJPRED, while SIMS and MINDIS violations were not different from each other. This leads to the overall ranking of $\{\text{SIMS}, \text{MINDIS}\} \gg \text{SUBJPRED}$. We conclude that SIMS qualifies as a hard constraint, as it leads to strong unacceptability, while SUBJPRED induces only mild unacceptability and thus should be classified as soft. The status of MINDIS is less clear, it seems to fall in between these two extremes (see Figure 4.7).

We also observed that the soft constraint SUBJPRED was subject to context effects (consider the increased effect of a SUBJPRED violation in the neutral context). On the other hand, SIMS, a hard constraint, was immune to context effects. This leads to the more general hypothesis that soft are constraints context-dependent, i.e., constraint violations are subject to context effects, while hard constraints are context-independent, i.e., immune to context effects (see Section 4.1.1). If correct, this hypothesis would provide us with an new diagnostic for the hard/soft distinction, in addition to constraint strength. Thus we can classify MINDIS as a soft constraint, as it is clearly subject to context effects, even though its constraint strength is in between that of the hard constraint SIMS and that of the soft constraint SUBJPRED.

This leads to the conclusion that the ranking of soft and hard constraints can be fairly similar, as for MINDIS and SIMS. In such a case, we cannot determine the type of a constraint solely based on the degree of unacceptability caused by its violation. Rather, other criteria such as context effects (or crosslinguistic effects, see Experiments 1–3) have to be taken into account to classify the constraint.

4.3.5.3. Constraint Interaction

The findings of Experiment 8 confirm another result from Chapter 3: constraint violations are cumulative, i.e., the degree of unacceptability increases with the number of violations. A cumulativity effect was obtained for both the null context condition and the neutral context condition (see Figure 4.8).

4.3.6. Conclusions

Experiment 8 extended the results from Experiment 7 by providing evidence for three constraints on gapping: MINDIS, SUBJPRED, and SIMS. It allowed us to classify MINDIS and SUBJPRED as soft constraints and SIMS as a hard constraint.

We also determined the interaction of gapping with context, investigating for four types of contexts: null context, neutral context, felicitous and infelicitous context. It was demonstrated that MINDIS and SUBJPRED are subject to context effects, while SIMS failed to show context effects. This led to the more general hypothesis that soft constraints are context-dependent, while hard constraints are context-independent. In the remainder of this chapter we

will investigate this hypothesis in more detail; Experiment 9 will deal with context effects on extraction, while Experiments 10–12 will investigate the interaction of word order and context.

Note that the present experiment led to the conclusion that soft and hard constraints can receive similar rankings (as was the case for MINDIS and SIMS. This indicates that a constraint cannot be classified as hard or soft based solely on its constraint rank. Rather, context effect and crosslinguistic effects have to be taken into account.

Finally, present experiment provided support for the cumulativity of constraint violations, thus extending the results on cumulativity already obtained in Experiments 4–6.

4.4. Experiment 9: Effect of Context on Extraction from Picture NPs

In Experiments 1–8 we provided evidence for the distinction between soft and hard constraint violations. We showed that soft violations cause only mild unacceptability, while hard violations lead to a high degree of unacceptability. Experiments 7 and 8 generated a new hypothesis regarding the soft/hard dichotomy: soft constraints are context-dependent, while hard constraints are context-independent.

The present experiment was designed to test this hypothesis with respect to the constraints investigated in Experiment 4. Recall that Experiment 4 dealt with extraction from picture NPs and showed that the constraints REFERENTIALITY (referentiality of the *wh*-phrase), DEFINITENESS (definiteness of the picture NP), and VERBCLASS (semantic class of the main verb) all have a weak, but significant influence on the acceptability of extraction. We therefore classified these constraints as soft. A set of hard constraints on extraction was also identified: INVERSION (inversion), RESUMPTIVE (resumptive pronouns), and AGREEMENT (subject-verb agreement). It was demonstrated that these constraints have a strong effect on acceptability.

4.4.1. Introduction

Definite noun phrases are context-dependent elements; they presuppose the existence of the object they refer to. One way of satisfying this presupposition is by providing a context that establishes the existence of the referent, e.g., using a deictic or an indefinite NP. In Experiment 4, we demonstrated that extraction from definite picture NPs is less acceptable than extraction from indefinite ones. It can be hypothesized that this definiteness effect is due to the context dependence of definites: a null context fails to provide an antecedent for the definite picture NP, which causes the reduced acceptability for extraction. Our prediction is that this effect should disappear in a context that is felicitous for definites, i.e., that establishes a referent which the definite NP can be bound to.

An example for such a context is given in (4.23c,d). The NP *this photograph* estab-

lishes a referent for the definite picture NP. In examples (4.23a,b), we give a neutral context that fails to provide such a referent, and thus should preserve the definiteness effect. The neutral context will function as a control condition in our experiment.

(4.23) a. Thomas seems to be very talented. Which friend has he taken a photograph of?
b. Thomas seems to be very talented. Which friend has he taken the photograph of?
c. Thomas has taken this photograph of a friend. Which friend has he taken a photograph of?
d. Thomas has taken this photograph of a friend. Which friend has he taken the photograph of?

Pesetsky (1987) deals with *wh*-extraction by making use of the notion of discourse linking: a *wh*-phrase is discourse linked if it refers to an object previously established in the discourse. Such an approach is useful in accounting for the referentiality effect on extraction of picture NPs demonstrated in Experiment 4: extraction is more acceptable if the extracted *wh*-phrase is referential (e.g., *which friend*), and less acceptable if it is non-referential like (e.g., *how many friends*). Pesetsky's (1987) account predicts that discourse linking is responsible for the referentiality effect; only referential NPs are inherently discourse linked. Hence the unacceptability of extracting a non-referential *wh*-phrase should disappear in a context where the *wh*-phrase is discourse linked. An example for such a context is given in (4.24d), where the non-referential phrase *how many friends* can be discourse linked to the phrase *some of his friends*. A similar context is provided for the referential NP *which friend* in (4.24c), so as to make the two cases comparable. The prediction is that there should be no difference in acceptability between (4.24c) and (4.24d), whereas the referentiality effect should be preserved in a neutral context such as the on in (4.24a,b).

(4.24) a. Thomas seems to be very talented. Which friend has he taken a photograph of?
b. Thomas seems to be very talented. How many friends has he taken a photograph of?
c. Thomas has taken a photograph of one of his friends. Which friend has he taken a photograph of?
d. Thomas has taken a photograph of some of his friends. How many friends has he taken a photograph of?

Finally, Experiment 4 demonstrated an effect of verb class on picture NP extraction. This effect is discussed by Diesing (1992), who claims that extraction from picture NPs is blocked if the matrix verb presupposes the existence of the picture NP. This is the case for [+EXISTENCE] verbs such as *tear up*. On the other hand, [−EXISTENCE] verbs such as *paint* or *take* do not carry this presupposition, and thus allow extraction.

Diesing (1992) assumes that picture NP extraction is possible even for a [+EXISTENCE] verb if the verb has a habitual reading, in which case no existential presup-

position is available. She gives examples like the one in (4.25d), where the context induces a habitual reading of *destroy*. The prediction is that there is no difference in acceptability between (4.25c) and (4.25d), while the effect of verb class is preserved in a neutral context like the one in (4.25a,b).

(4.25) a. Thomas seems to be very talented. Which friend has he taken a photograph of?
b. Thomas seems to be very angry. Which friend has he destroyed a photograph of?
c. Thomas takes a photograph of one of his friends every week. Which friend has he taken a photograph of this week?
d. Thomas destroys a photograph of one of his friends every week. Which friend has he destroyed a photograph of this week?

While context effects are predicted for soft constraints, hard constraints should be context-independent. To test this hypothesis, the present experiment included the hard constraints on inversion, resumptive pronouns, and agreement that were shown to have an effect on extraction in Experiment 4. It is not clear what felicitous contexts for these constraints could look like—which is of course why we hypothesize that hard constraints are context-independent. It makes sense, however, to include contexts like the ones used for the soft constraints as a control condition. This allows us to show that it is not the context as such that improves acceptability, but the interaction between the context and a specific constraint violation. The contexts used for hard violations were ones employed for the REF violation. As an example, consider the INV constraint, presented in a neutral context in (4.26a,b) and in a felicitous context in (4.26c,d). The same contexts were used for the AGR and RES violation.

(4.26) a. Thomas seems to be very talented. Which friend has he taken a photograph of?
b. Thomas seems to be very talented. Which friend he has taken a photograph of?
c. Thomas has taken a photograph of one of his friends. Which friend has he taken a photograph of?
d. Thomas has taken a photograph of one of his friends. Which friend he has taken a photograph of?

4.4.2. Predictions

The present experiment used the same constraints that were already shown to have an effect on extraction from picture NPs in Experiment 4. Thus we expect that the effects of constraint violations found in the earlier experiment will be replicated.

Our hypothesis is that soft constraints are context-dependent, while hard constraints are context-independent. Hence we predict that a violation of a soft constraint like DEF, REF, and VERB should disappear in a felicitous context, but should be preserved in a neutral context, i.e., there should be an interaction of constraint violation and context.

For the hard constraints INV, AGR, and RES, no such interaction is predicted. Hard constraint violations should be equally unacceptable both in a neutral and in a felicitous context.

4.4.3. Method

4.4.3.1. Subjects

Thirty-one native speakers of English from the same population as in Experiment 4 participated in the experiment. None of the subjects had previously participated in Experiment 4.

The data of a one subject were eliminated after an inspection of the responses showed that he had not completed the task adequately.

This left 30 subjects for analysis. Of these, 15 subjects were male, 15 female; four subjects were left-handed, 26 right-handed. The age of the subjects ranged from 17 to 67 years, the mean was 28.8 years.

4.4.3.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials The experiment included a total of six subdesigns, three for soft constraints and three for hard constraints. Each of the subdesigns crossed the factors constraint violation (*Viol*) and context (*Con*).

The soft constraints were the ones already investigated in Experiment 4: definiteness (DEF), referentiality (REF), and verb class (VERB). The factor *Viol* had two levels (constraint violation or no violation, see (a) and (b) examples in (4.23)–(4.25)). Two contexts were used (neutral context and felicitous context, see (a) and (c) examples in (4.23)–(4.25)). This yielded three subdesigns with $Viol \times Con = 2 \times 2 = 4$ cells. Duplicate cells were presented only once (the neutral context was the same for all constraints), which reduced the number of cells to ten.

The hard constraint were the same as in Experiment 4: inversion (INV), resumptive pronouns (RES), and agreement (AGR). Again, the factor *Viol* had two levels (constraint violation or no violation, as illustrated in examples (3.33)–(3.35)). The two levels for the factor context were the same ones as for the REF constraint (neutral and felicitous, see (a) and (c) examples in (4.26)). This yielded three subdesigns with $Viol \times Con = 2 \times 2 = 4$ cells. Duplicate cells were presented only once (the no violation condition was the same as in the first subexperiment), which reduced the number of cells to six. Four lexicalizations were used for each cell in each subexperiment, which resulted in a total of 64 stimuli.

A set of 16 fillers was used, designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

The lexicalizations were matched for frequency using the same procedure as in Experiment 4.

4.4.3.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions We used an English version of the instructions in Experiment 1. Where contextualized stimuli were presented, subjects were told that each sentence would be presented in context, defined as a single sentence preceding the target sentence. Subjects were instructed to judge the acceptability of the target sentence, and to take the context into account in their judgments. The task was illustrated by examples.

Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

Four test sets were used: each test set contained one lexicalization for each of the 16 cells in the design. Lexicalizations were assigned to test sets using a Latin square covering the full set of items.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 32 test items: 16 experimental items and 16 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each experimental subject was randomly assigned to one of the test sets.

4.4.4. Results

The data were normalized as in Experiment 1. Due to the design of the present experiment, we could not carry out an omnibus ANOVA: the data for the six constraint violations overlapped (both for soft and hard violations), as we presented the null violation condition only once. Hence we carried out a series of planned comparisons and adjusted the significance level using the Bonferroni method. As six comparisons were carried out in total, so we set $p = .0083$.

Soft Constraints Figure 4.9 graphs the interaction between violation and context for the three soft constraints. We conducted a planned comparison for each of the constraints, using an ANOVA with the factors constraint violation and context.

For definiteness and referentiality, no significant main effects or interactions were found. Note however, that there was a tendency in the predicted direction for both constraints: a violation of DEF or REF is less acceptable than no violation, and this difference disappears in a felicitous context (see Figure 4.9a,b). For verb class, we found a main effect of violation (significant by subjects only, $F_1(1,29) = 9.015$, $p = .005$; $F_2(1,3) = 1.040$, $p = .383$). A main effect of context was also found (significant by subjects only, $F_1(1,29) = 11.559$, $p = .002$; $F_2(1,3) = 1.232$, $p = .348$). There was no interaction of violation and context, i.e., context did

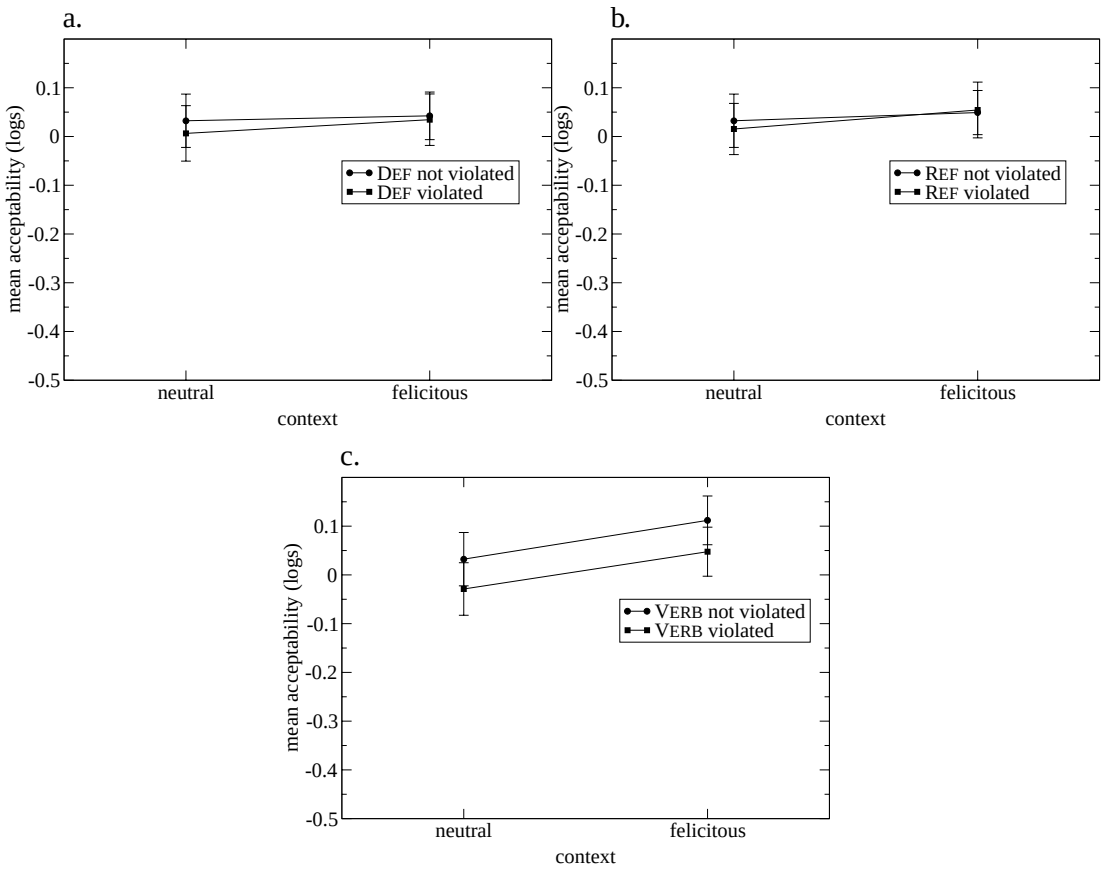

Figure 4.9: Context effects for DEF, REF, and VERB, single violations (Experiment 9)

not reduce the seriousness of a VERB violation, even though it increased acceptability overall, as illustrated by Figure 4.9c.

Hard Constraints Figure 4.10 graphs the interaction between violation and context for the tree hard constraints. Again, we conducted a planned comparison for each of the constraints, using an ANOVA with the factors constraint violation and context.

For inversion, the main effect of violation was marginal by subjects ($F_1(1,29) = 6.751$, $p = .015$; $F_2(1,3) = 1.700$, $p = .283$). Main effects of violation were found for agreement (significant by subjects and marginal by items, $F_1(1,29) = 39.459$, $p < .0005$; $F_2(1,3) = 28.210$, $p = .013$) and for resumptive pronouns ($F_1(1,29) = 46.612$, $p < .0005$; $F_2(1,3) = 67.772$, $p = .004$). None of the constraints exhibited a main effect of context or an interaction of context and violation, which confirms our hypothesis that context fails to influence on hard violations.

4.4.5. Discussion

We predicted that the constraint violations investigated in Experiment 4 would also have an effect on extraction from picture NPs in the present experiment. This was the case for the hard

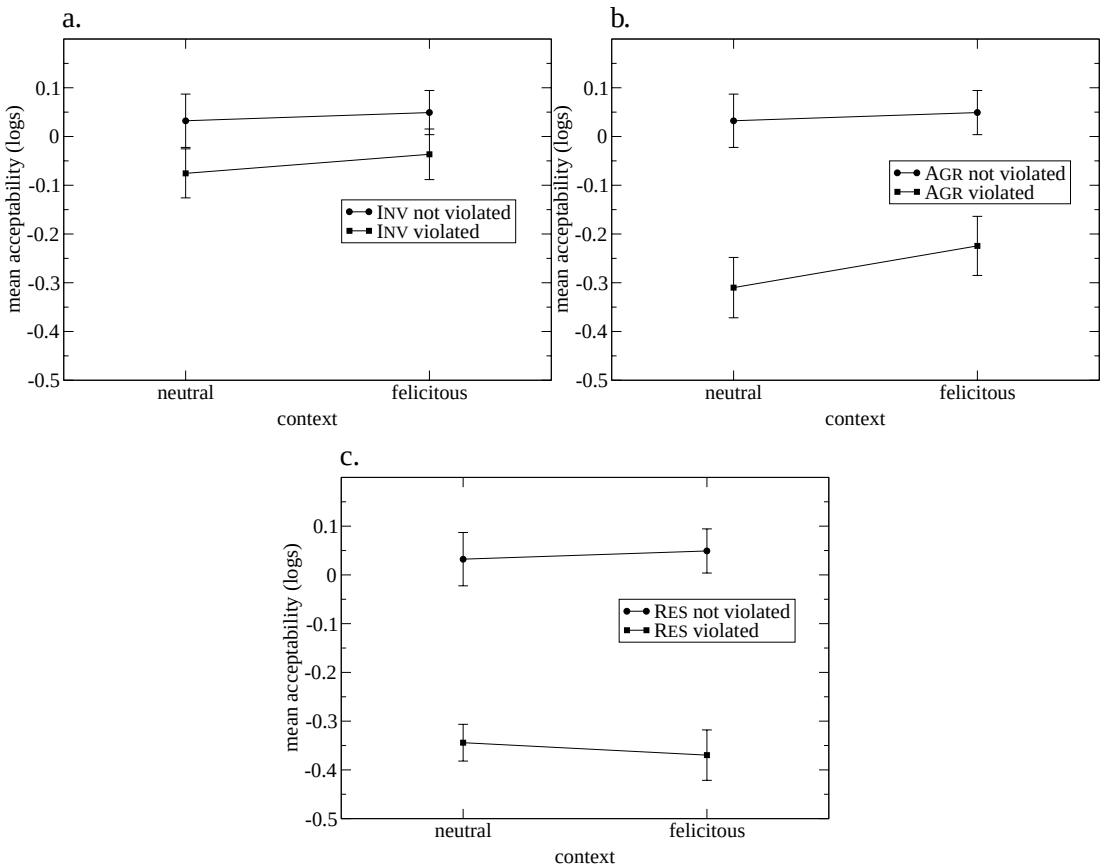

Figure 4.10: Context effects for INV, AGR, and RES, single violations (Experiment 9)

violations (INV, AGR, and RES), which triggered a significant decrease in acceptability.

For soft constraints, we replicated the effect of VERB on extraction from picture NPs. However, no effects of DEF and RES were obtained. This might mean that these effects are too small to be detected in the present experiment, where only single constraint violations were tested. Note that also in Experiment 4, the effects of DEF and REF were (non-significantly) smaller than the effect of VERB (see Figure 3.7).

The present experiment was designed to investigate contextual effects on hard and soft constraint violations. The prediction was that soft violations will disappear in a felicitous context, but will be preserved in a neutral context. However, we failed to find an effect of constraint violation for DEF and REF. This could mean that these violations disappear in both contexts, i.e., even in a neutral context. On the other hand, a tendency in the predicted direction was observed (see Figure 4.9a,b): the difference between violation and non-violation seems to disappear in the felicitous context. The fact that this difference failed to reach significance might be due to the small size of the effect.

For VERB (see Figure 4.9c), we also failed to find the predicted interaction of constraint violation and context, though there was a main effect of violation. However, a main

effect of context was obtained: a habitual context like in (4.25c,d) is more acceptable than a neutral context like in (4.25a,b). The claim in the theoretical literature (see Diesing 1992) is that extraction from picture NPs is acceptable for a [+EXISTENCE] verb like *destroy* if the verb has a habitual reading. The experimental results confirm that a habitual context improves the acceptability of extraction for [+EXISTENCE] verbs (as claimed in the literature), but also show that the same improvement occurs for [−EXISTENCE] verbs.

For the hard constraints INV, AGR, and RES we expected the absence of an interaction between violation and context. This prediction was borne out—there was no difference between the acceptability of hard constraint violations in a neutral and in a felicitous context.

4.4.6. Conclusions

This experiment focused on the hypothesis that context effects can serve as a diagnostic for the soft/hard distinction of constraint violations. As predicted, context effects were absent for hard constraints. The evidence regarding context effects for soft constraints was less clear. Instead of obtaining the predicted interaction between violation and context, we found that certain soft violations (DEF and REF) disappear completely in context (even in a neutral context). For VERB violations, on the other hand, main effects of violation and context were observed, but no interaction.

While this evidence is compatible with our general hypothesis that soft constraints, but not hard constraints, are context-dependent, further evidence from other phenomena is required to support this hypothesis. In Experiments 10–12 we will therefore return to a phenomenon that was already investigated in Experiment 6: word order. We will first extend our results on word order preferences in German, and then provide data on word order in Greek, thus adding a crosslinguistic dimension to our results. We will also conduct an experiment using spoken stimuli, which allows us to test the interaction of syntactic and phonological constraints, broadening the range of evidence considered.

4.5. Experiment 10: Effect of Case, Pronominalization, Verb Position, and Context on Word Order

In Experiment 6 we investigated gradient acceptability for complement ordering in the subordinate clause in German. We found evidence for three linear precedence constraints: NOMALIGN, which states that nominative NPs have to precede non-nominative NPs, DATALIGN, which requires that dative NPs precede accusative NPs, and PROALIGN, specifying that pronouns have to precede full NPs. Experiment 6 also demonstrated that NOMALIGN and PROALIGN outrank DATALIGN; NOMALIGN and PROALIGN, on the other hand, were found to be ranked equally. The present experiment extends the results of Experiment 6. It is

designed to further investigate PROALIGN and NOMALIGN, but also includes the additional constraints VERBFINAL (the verb has to succeed any other constituent) and GROUNDALIGN (ground constituents have to be sentence peripheral).

The main aim of this experiment is to supply further evidence for the hypothesis that soft constraints are context-dependent, while hard constraints are context-independent (see the discussion in Section 4.1.1). This hypothesis is based on the results on gapping and extraction obtained in Experiments 7–9. The present experiment extends the investigation of context effect to word order preferences. It will also supply additional data on constraint ranking and constraint interaction.

4.5.1. Introduction

For its investigation of complement order in German, the present experiment used subordinate clause stimuli (see Section 3.7.1 for a brief motivation). While Experiment 6 dealt with ditransitive verbs, the present experiment investigated transitive verbs, limiting the range of complement orders to subject before object and object before subject (only accusative objects were included). To complement the results of Experiment 6, the present study also manipulated verb order by investigating verb final and verb initial stimuli. Pronominalization was again included as a factor.[6]

The combination of two complement orders and two verb orders yields a total of four word orders, illustrated by the examples in (4.27). As was discusses already in Section 3.7.1, subordinate clauses in German require verb final order (see (4.27a,b)). Verb initial orders (see (4.27c,d)) are expected to give rise to strong unacceptability.

(4.27) a. **SOV:** Maria glaubt, dass der Vater den Wagen kauft.
Maria-NOM believes that the father-NOM the car-ACC buys
"Maria believes that the father will buy the car."

b. **OSV:** Maria glaubt, dass den Wagen der Vater kauft.

c. * **VSO:** Maria glaubt, dass kauft der Vater den Wagen.

d. * **VOS:** Maria glaubt, dass kauft den Wagen der Vater.

To examine the influence of pronominalization on word order, the experiment included sentences where none of the NPs was pronominalized (see (4.27)), but also sentences where the subject, object, or both the subject and the object were pronominalized (see (4.28)).

(4.28) a. Maria glaubt, dass er den Wagen kauft.
Maria-NOM believes that he-NOM the car-ACC buys
"Maria believes that he will buy the car."

[6] In contrast to Experiment 6, the present study used inanimate accusative NPs. This move can be justified by the fact that Experiment 6 (using animate accusative NPs) yielded the same acceptability ranking as Pechmann et al.'s (1994) study (using inanimate accusative NPs), indicating the weak influence of animacy on word order preferences in German.

b. Maria glaubt, dass der Vater ihn kauft.
Maria-NOM believes that the father-NOM it-ACC buys
"Maria believes that the father will buy it."

c. Maria glaubt, dass er ihn kauft.
Maria-NOM believes that he-NOM it-ACC buys
"Maria believes that he will buy it."

Information Structure figures as a determinant of complement order in the accounts of Choi (1996), Jacobs (1988), Müller (1999), and Uszkoreit (1987). Information structural effects can be studied by embedding the sentence in a question context: the *wh*-phrase marks the focussed constituent, while the other constituents are non-focussed, or ground (Vallduví 1992). (The information structural constraint GROUNDALIGN is discussed in Section 3.7.1; for details on the information structural framework we assume see Section 4.6.1.)

The following contexts were used in the experiment:

(4.29) a. **Null**

b. **All Focus:** Was gibt's neues?
"What's new?"

c. **S Focus:** Wer kauft den Wagen?
"Who will buy the car?"

d. **O Focus:** Was kauft der Vater?
"What will the father buy?"

A null context condition was included as a control, allowing us to study how subjects react in the absence of any contextual information.

4.5.2. Predictions

4.5.2.1. Constraints out of Context

In Experiment 6 we demonstrated that violations of the constraints NOMALIGN and PROALIGN lead to a significant reduction in acceptability in non-contextualized stimuli. In the present experiment, we expect to replicate these effects: NOMALIGN predicts that orders where the subject precedes the object are more acceptable than orders where the object precedes the subject, while PROALIGN predicts that orders where a pronoun precedes a full NP are more acceptable than orders were a full NP precedes a pronoun. (Note that the present experiment only deals with single violations of NOMALIGN and PROALIGN, as the materials use transitive verbs. Experiment 6 used ditransitive verbs, and thus could investigate multiple violations of these constraints.)

In addition to the effects of NOMALIGN and PROALIGN, we expect to find an effect of VERBFINAL, which predicts that verb final orders are more acceptable than verb initial orders.

4.5.2.2. Constraints in Context

NOMALIGN is classified as a soft constraint by Müller (1999) (a markedness constraint in his terminology). The results of Experiment 6 were consistent with this; we found that NOMALIGN violations lead to only mild unacceptability. Müller (1999) classifies PROALIGN as a hard constraint; a hypothesis that could not be confirmed by Experiment 6, where PROALIGN and NOMALIGN were found to have the same rank. We concluded that both are soft constraints.

Hence we now predict that both PROALIGN and NOMALIGN will be subject to context effects, i.e., the effects of these constraints will be weaker in certain contexts, or disappear completely. (Recall that Experiments 7 and 8 lead to the hypothesis that soft constraints are context-dependent, while hard constraints are immune to context effects.)

The constraint GROUNDALIGN requires that ground (non-focussed) constituents are peripheral, i.e., occur sentence initially or sentence finally. This means that the acceptability of SOV should be reduced in the S focus context, where this order violates GROUNDALIGN (as the ground object is non-peripheral). In the O focus context, on the other hand, we expect OSV to be less acceptable, as the ground subject is non-peripheral in this order, thus violating GROUNDALIGN. If GROUNDALIGN turns out to be a soft constraints, we expect it to be subject to context effects.

Intuitively, the constraint VERBFINAL that regulates the order of the verb in the subordinate clause is very strong; we expect VERBFINAL to be a hard constraint, i.e., it should be context-independent.

4.5.2.3. Constraint Ranking

In Experiment 6 we found that NOMALIGN and PROALIGN were ranked equally. We predict that this ranking will be replicated in the present experiment. As for the constraint GROUNDALIGN, Müller (1999) assumes the ranking NOMALIGN $\gg$ GROUNDALIGN, which predicts that a NOMALIGN violation should lead to greater unacceptability than a GROUNDALIGN violation (under the operational definition of constraint ranking adopted in Section 3.1.2). The constraint VERBFINAL (not explicit dealt with by Müller) is expected to be a hard constraint, i.e., it should outrank all the other constraints.

As in previous experiments, we will test predictions on constraint rankings by carrying out planned comparisons involving single constraint violations.

4.5.2.4. Constraint Interaction

In Experiment 6, we found evidence for the fact that constraint violations are cumulative, i.e., that the degree of unacceptability of a stimulus increases with the number of violations incurred. This effect occurred both for multiple violations of the same constraint and for multiple violation of different constraints.

The present experiment contains sentences that incur between zero and three violations of the constraints NOMALIGN, PROALIGN, and VERBFINAL. We expect these violations to be cumulative, a prediction that can be put to the test by carrying out a set of planned comparisons involving multiple constraint violations. (Note, however, that the cumulativity of violations of GROUNDALIGN cannot be tested directly, as this constraint is expected to interact with NOMALIGN and PROALIGN.)

4.5.3. Method

4.5.3.1. Subjects

Fifty-eight native speakers of German from the same population as in Experiment 1 participated in the experiment. None of the subjects had previously participated in Experiment 6.

The data of three subjects were excluded because they were bilingual (by self-assessment). The data of another two subjects were excluded because they were linguists (by self-assessment). The data of a two subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 51 subjects for analysis. Of these, 37 subjects were male, 14 female; three subjects were left-handed, 48 right-handed. The age of the subjects ranged from 19 to 45 years, the mean was 28.7 years.

4.5.3.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials A factorial design was used that crossed the factors verb order (*Vord*), complement order (*Cord*), pronominalization (*Pro*), and context (*Con*). The factor *Con* had four levels: null context, all focus, S focus, and O focus, as illustrated in (4.29). The factor *Vord* had two levels: verb final (see (4.27a,b)) and verb initial (see (4.27c,d)). The two levels of *Cord* were subject before object and object before subject, as in (4.27a,c) and (4.27b,d). In the null context condition, the factor *Pro* had four levels, viz., both S and O full NPs, S pronoun and O full NP, S full NP and O pronoun, and both S and O pronouns (see (4.28)). In the context condition, *Pro* only had two levels, viz., no pronoun and pronoun. In the all focus and S focus contexts, the object was pronominalized, while in the O focus context, the subject was pronominalized. This design ensures that the pronoun is interpreted as ground and hence is unstressed (as the sentential stress has to fall on the focussed constituent). We are only interested in the syntactic behavior of weak (i.e., unstressed) pronouns; strong (i.e., stressed) pronouns are subject to different syntactic constraints (Müller 1999).

This yielded a total of $Vord \times Cord \times Pro = 2 \times 2 \times 4 = 16$ cells for the null context condition, and $Vord \times Cord \times Pro \times Con = 2 \times 2 \times 2 \times 3 = 24$ cells for the context condition. Eight lexicalizations per cell were used, which resulted in a total of 320 stimuli.

A set of 24 fillers was used in the null context condition; 16 fillers were employed in the context condition. The fillers were designed to cover the whole acceptability range. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

To control for possible effects from lexical frequency, the lexicalizations for subject, object, and verb were matched for frequency. Frequency counts for the verbs and the head nouns were obtained from a lemmatized version of the Frankfurter Rundschau corpus (40 million words of newspaper text) and the average frequencies were computed for subject, object, and verb lexicalizations. An ANOVA confirmed that these average frequencies were not significantly different from each other.

4.5.3.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions We used the same instructions as in Experiment 1. Where contextualized stimuli were presented, subjects were told that each sentence would be presented in context, defined as a single sentence preceding the target sentence. Subjects were instructed to judge the acceptability of the target sentence, and to take the context into account in their judgments. The task was illustrated by examples.

Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

A between subjects design was used to administer the factor *Con*: subjects in Group A judged non-contextualized stimuli, while subjects in Group B judged contextualized stimuli. The factors *Vord*, *Cord*, and *Pro* were administered within subjects.

For both groups, eight test sets were generated: for Group A, each test set contained one lexicalization for each of the 16 cells in the first subdesign. For Group B, each test set contained one lexicalization for each of the 24 cells in the second subdesign. Lexicalizations were assigned to test sets using Latin squares. Two separate Latin squares were applied: one for the null context condition and one for the context condition.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 40 test items: 16 experimental items and 24 fillers in Group A, and 24 experimental items and 16 fillers in Group B. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to a group and a test set; 20 subjects were assigned to group A,

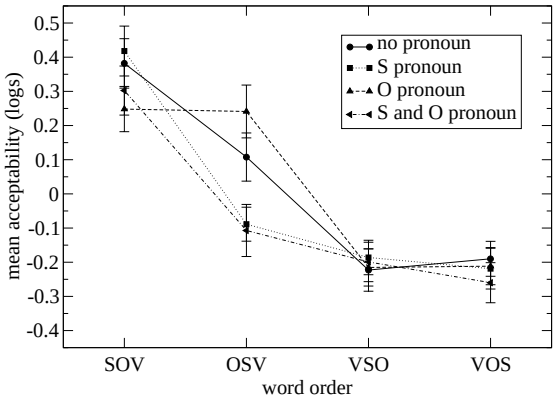

Figure 4.11: Interaction for word order and pronominalization, null context condition (Experiment 10)

and 23 to group B. Instructions, examples, training items, and fillers were adapted for Group B to take context into account.

4.5.4. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

In discussing the results, we make use of the following abbreviations: SO for subject before object, OS for object before subject, XV for verb final, VX for verb initial. The indices "pro" and "full" indicate pronouns and full NPs, respectively. For instance, $VS_{full}O_{pro}$ stands for a VSO order where the subject is a full NP and the object is a pronoun. We leave out the subscript when we disregard the distinction between full and pronominalized NPs.

4.5.4.1. Constraints out of Context

Figure 4.11 graphs the average judgments for each word order in the null context condition. An ANOVA for the null context condition revealed a highly significant main effect of *Vord* (verb order) ($F_1(1,19) = 56.911$, $p < .0005$; $F_2(1,7) = 621.924$, $p < .0005$): XV orders (mean $= .1879$) were more acceptable than VX orders (mean $= -.2129$). A highly significant main effect of *Cord* (complement order) was also obtained ($F_1(1,19) = 26.966$, $p < .0005$; $F_2(1,7) = 72.610$, $p < .0005$): SO orders (mean $= .0659$) were more acceptable than OS orders (mean $= -.0909$). The main effect of *Pro* (pronominalization) was significant by subjects only ($F_1(3,57) = 5.150$, $p = .003$; $F_2(3,21) = .647$, $p = .593$).

The ANOVA also revealed a significant interaction of *Cord* and *Pro* ($F_1(3,57) = 13.026$, $p < .0005$; $F_2(3,21) = 4.663$, $p = .012$). This indicates that pronominalization has an influence on complement order preference. We also found interactions of *Cord* and *Vord* ($F_1(1,19) = 47.437$, $p < .0005$; $F_2(1,7) = 17.148$, $p = .004$) and of *Vord* and *Pro* (signifi-

cant by subjects only, $F_1(3,57) = 4.223$, $p = .009$; $F_2(3,21) = 1.107$, $p = .368$). A three-way interaction *Vord/Cord/Pro* was also present (significant by subjects only, $F_1(3,57) = 7.415$, $p = .009$; $F_2(3,21) = 1.900$, $p = .161$). The meaning of the interactions involving *Vord* becomes clear from Figure 4.11: the effect of NOMALIGN and PROALIGN is limited to verb final orders; all verb initial orders are highly unacceptable, and only a very small influence of complement order and pronominalization is observed.

A post-hoc Tukey test was carried out to further investigate the interaction of *Cord* and *Pro*. For stimuli with two full NPs, it was found that $S_{full}O_{full}$ was more acceptable than $O_{full}S_{full}$ (by subjects only, $\alpha < .01$), in line with the predictions of NOMALIGN. For the stimuli with one pronominalized NP, $S_{pro}O_{full}$ (which satisfies NOMALIGN and PROALIGN) was more acceptable than $O_{pro}S_{full}$ (which violates NOMALIGN but satisfies PROALIGN) (by subjects only, $\alpha < .05$). We also found that $S_{pro}O_{full}$ was more acceptable than $S_{full}O_{pro}$ (which violates PROALIGN but satisfies NOMALIGN) (by subjects only, $\alpha < .05$). $S_{full}O_{pro}$ and $O_{pro}S_{full}$, on the other hand, were not significantly different. Furthermore, $O_{full}S_{pro}$ (which violates both PROALIGN and NOMALIGN) was less acceptable than $S_{pro}O_{full}$ ($\alpha < .01$), $S_{full}O_{pro}$ (by subjects only, $\alpha < .01$), and $O_{full}S_{pro}$ (by subjects only, $\alpha < .01$). For the stimuli with two pronominalized NPs, it was found that $S_{pro}O_{pro}$ was more acceptable than $O_{pro}S_{pro}$ ($\alpha < .01$), in line with the predictions of NOMALIGN.

4.5.4.2. Constraints in Context

Figure 4.12 graphs the average judgments for each context. An ANOVA for the context condition confirmed the main effect of verb order found in the null context condition ($F_1(1,30) = 121.507$, $p < .0005$; $F_2(1,7) = 225.903$, $p < .0005$): XV orders (mean $= .2519$) were more acceptable than VX orders (mean $= -.1973$). The main effect of complement order could also be replicated ($F_1(1,30) = 40.275$, $p < .0005$; $F_2(1,7) = 15.359$, $p = .006$): SO orders (mean $= .0785$) were more acceptable than OS orders (mean $= -.0239$). A highly significant main effect of *Con* (context) was also present ($F_1(2,60) = 28.953$, $p < .0005$; $F_2(2,14) = 54.056$, $p < .0005$), as well as a weak effect of *Pro* ($F_1(2,60) = 5.564$, $p = .025$; $F_2(2,14) = 1.511$, $p = .259$).

The ANOVA uncovered an interaction of *Cord* and context, significant by subjects and marginal by items ($F_1(2,60) = 6.016$, $p = .004$; $F_2(2,14) = 3.076$, $p = .078$), which confirms that Information Structure (manipulated by context) has an influence on complement order preferences. We also found a marginal interaction of *Cord* and *Pro* ($F_1(1,30) = 4.025$, $p = .054$; $F_2(1,7) = 3.634$, $p = .098$) and a highly significant interaction of *Pro* and context ($F_1(2,60) = 11.864$, $p < .0005$; $F_2(2,14) = 16.07$, $p < .0005$). Recall that our materials were designed such that in all focus and S focus contexts, the object was pronominalized, while in an O focus context, the subject was pronominalized. This means that the *Cord/Pro* and *Pro/Con* interactions are only meaningful with respect to the three-way interaction *Cord/Pro/Con*,

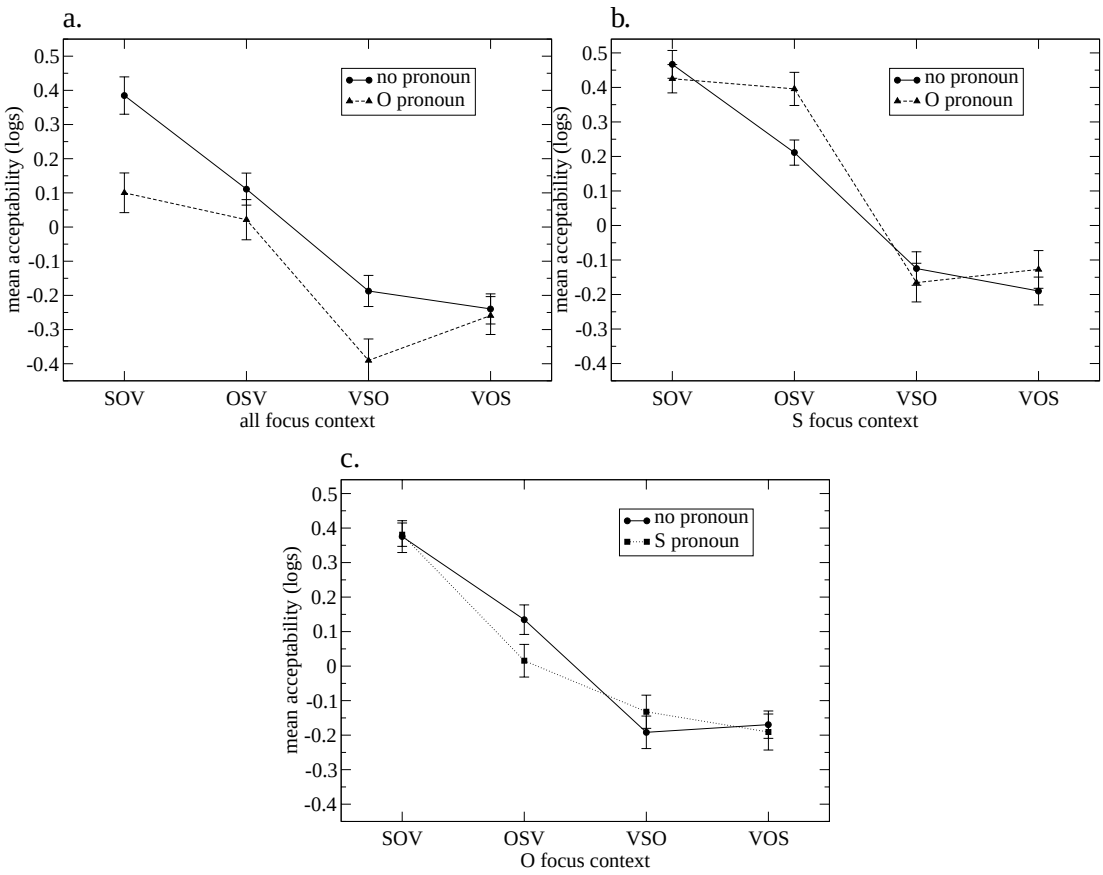

Figure 4.12: Interaction for word order and pronominalization, context condition (Experiment 10)

which was also significant ($F_1(2,60) = 19.718$, $p < .0005$; $F_2(2,14) = 7.73$, $p = .005$). This interaction demonstrates that the ordering of pronouns is subject to contextual effects (see results of the post-hoc test below). The ANOVA also showed an interaction of *Vord* and *Cord* ($F_1(1,30) = 50.960$, $p < .0005$; $F_2(1,7) = 7.221$, $p = .031$) and of *Vord* and context ($F_1(2,60) = 10.589$, $p < .0005$; $F_2(2,14) = 11.945$, $p = .001$). The meaning of the *Vord*/*Cord* interaction (see Figure 4.12) is the same as in the null context: the effect of NOMALIGN is limited to verb final orders; only a small effect of complement order seems to occur in verb initial orders.

To determine the effect of the constraint GROUNDALIGN, a Tukey test was conducted on the interaction of *Cord* and context. The Tukey results show that SO (which satisfies NOMALIGN) is more acceptable than OS (which violates NOMALIGN) in the all focus context (by subjects only, $\alpha < .05$), in the S focus context (by subjects only, $\alpha < .01$), and in the O focus context ($\alpha < .01$). We also found that OS was more acceptable in the S focus context than in the O focus context ($\alpha < .01$). This is because OS incurs a violation of GROUNDALIGN in the O focus context, but not in the S focus context (recall that GROUNDALIGN requires that ground

constituents have to be peripheral). The acceptability of SO, on the other hand, was the same in the S focus context and the O focus context, contrary to the predictions of GROUNDALIGN, which favors OS in S focus and SO in O focus. (In our discussion, we will disregard verb initial orders due to their general low acceptability.)

Another Tukey test was carried out to investigate the *Cord*/*Pro*/*Con* interaction (see also Figure 4.12). This test allows us to establish context effect for the constraints PROALIGN and NOMALIGN. In the all focus context, we found that $S_{full}O_{full}$ was more acceptable than $O_{full}S_{full}$ (by subjects, $\alpha < .01$, and by items, $\alpha < .05$), $S_{full}O_{pro}$ ($\alpha < .01$), and $O_{pro}S_{full}$ ($\alpha < .01$). This can be explained by the fact that $O_{full}S_{full}$, $S_{full}O_{pro}$, and $O_{pro}S_{full}$ all incur a violation of either PROALIGN or NOMALIGN, while $S_{full}O_{full}$ does not incur any violations. In the S focus context, $S_{full}O_{full}$ was more acceptable than $O_{full}S_{full}$ (by subjects, $\alpha < .01$, and by items, $\alpha < .05$), in line with the predictions of NOMALIGN. Furthermore, $S_{full}O_{pro}$ and $O_{pro}S_{full}$ were more acceptable than $O_{full}S_{full}$ (by subjects only, $\alpha < .01$ in both cases). However, $S_{full}O_{full}$, $S_{full}O_{pro}$, and $O_{pro}S_{full}$ failed to differ in acceptability, although $S_{full}O_{pro}$ violates PROALIGN, and $O_{pro}S_{full}$ violates NOMALIGN, while $S_{full}O_{full}$ does not incur any violations. In the O focus context, we found that $S_{full}O_{full}$ was more acceptable than $O_{full}S_{full}$ (by subjects only, $\alpha < .01$) and $O_{full}S_{pro}$ ($\alpha < .01$). Furthermore, $S_{pro}O_{full}$ was more acceptable than $O_{full}S_{full}$ and $O_{full}S_{pro}$ ($\alpha < .01$ in both cases). $O_{full}S_{full}$ and $O_{full}S_{pro}$ did not differ in acceptability, even though $O_{full}S_{pro}$ violates PROALIGN, while $O_{full}S_{full}$ does not.

4.5.4.3. Constraint Ranking

A separate analysis was conducted to determine constraint rankings. As in our previous experiment on word order (Experiment 6), we carried out a series of planned comparisons to compare the degree of unacceptability caused by single constraint violations. This analysis could only be applied to the null context condition; in the context condition an unexpected interaction of the constraints PROALIGN and context was found, which makes it impossible to directly compare the effect of single constraint violations.

Two sets of comparisons were carried out: one for stimuli involving two full NPs or two pronouns, and one for stimuli involving one full NP and one pronoun. The first data set allows to compare single violations of NOMALIGN with single violations of VERBALIGN, while the second data sets allows comparisons of single violations of NOMALIGN, PROALIGN, and VERBALIGN. Single violations for both data sets are graphed in Figure 4.13. Three planned comparisons were carried out on the second data set, hence we set $p = .0167$ (Bonferroni adjustment).

For the stimuli with two full NPs or two pronouns, we found that a violation of VERBFINAL (mean $= -.2110$) was significantly more serious than a violation of NOMALIGN (mean $= .0004$) ($F_1(1,19) = 11.960$, $p = .003$; $F_2(1,7) = 21.234$, $p = .002$).

For the data set with one full NP and one pronoun, we found that a violation of

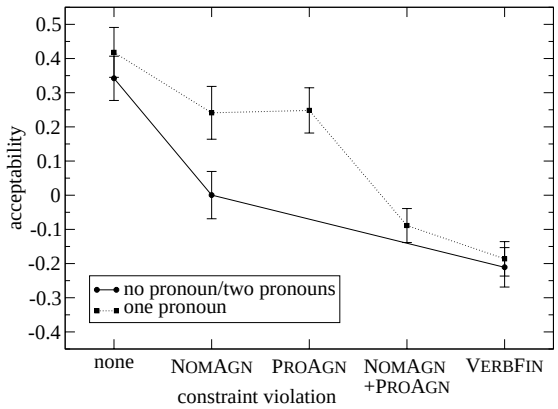

Figure 4.13: Constraint ranking, single violations (Experiment 10)

VERBFINAL (mean $= -.1861$) was more serious than a violations of NOMALIGN (mean $=$ .2412) ($F_1(1,19) = 29.076$, $p < .0005$; $F_2(1,7) = 33.871$, $p = .001$). Also, a violation of VERBFINAL was more serious than a violation of PROALIGN (mean $= .2482$) ($F_1(1,19) = 39.445$, $p < .0005$; $F_2(1,7) = 46.865$, $p < .0005$), while NOMALIGN and PROALIGN violations were not significantly different.

4.5.4.4. Constraint Interaction

A further set of planned comparisons on the data from the null context condition was conducted to determine if constraint violations were cumulative (in line with the cumulativity effect found in Experiments 4–6 and 8). For the subset of the data that contained two full NPs or two pronouns, we compared multiple violations of NOMALIGN and VERBALIGN by computing mean acceptability scores for stimuli with zero violations (two sentence types), one violation (four sentence types), and two violations (two sentence type). For the data set with one full NP and one pronoun, we compared multiple violations of NOMALIGN, PROALIGN, and VERBALIGN by computing mean acceptability scores for stimuli with zero violations (one sentence type), one violation (three sentence types), two violations (three sentence types), and three violations (one sentence type). The average judgments for both subsets are graphed in Figure 4.14. A Bonferroni adjustment was carried out in each case, leading to a significance level of $p = .025$ and $p = .0167$, respectively.

For the first data set, we found that zero constraint violations (mean $= .3421$) were more acceptable than a single violation (mean $= -.1053$) ($F_1(1,19) = 67.368$, $p < .0005$; $F_2(1,7) = 62.927$, $p < .0005$), which in turn was more acceptable than a double violation (mean $= -.2250$) ($F_1(1,19) = 10.595$, $p = .004$; $F_2(1,7) = 9.692$, $p = .017$).

For the second data set, we found that zero violations (mean $= .4180$) were better than a single violation (mean $= .1011$) ($F_1(1,19) = 45.739$, $p < .0005$; $F_2(1,7) = 13.633$, $p =$

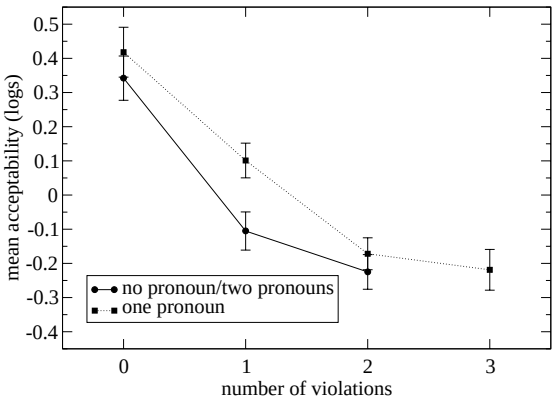

Figure 4.14: Cumulativity of constraint violations (Experiment 10)

.008). Also a single violation was better than a double violation (mean $= -.1718$) ($F_1(1,19) = 57.841$, $p < .0005$; $F_2(1,7) = 57.898$, $p < .0005$). The difference between a double violation and a triple violation (mean $= -.2188$), however, failed to reach significance.

Overall, these results confirm the cumulativity effect that was already established in Experiments 4–6 and 8.

Furthermore, we tested for ganging up effects in the second data set, i.e., the one involving one full NP and one pronoun. (Recall that ganging up effects were already found in Experiments 4 and 5.) We conducted a post-hoc test to determine if a combined violation of NOMALIGN and PROALIGN is as serious as a single violation of VERBALIGN (see Figure 10). As in Experiments 4 and 5, the post-hoc test employed the same significance level as the planned comparisons used to determine constraint ranks, i.e., $p = .0167$. It was found that the difference between a combined violation of NOMALIGN and PROALIGN (mean $= -.0887$) and a single violation of VERBALIGN (mean $= -.1861$) was only marginal by subjects ($F_1(1,19) = 6.559$, $p = .019$; $F_2(1,7) = 6.351$, $p = .040$).

4.5.5. Discussion

4.5.5.1. Constraints out of Context

The experimental findings for the null context provided clear support for the ordering constraint NOMALIGN, which requires nominative to precede accusative, in line with the results of Experiment 6. In addition, we showed that the constraint VERBFINAL correctly describes the verb position in subordinate clauses: there was a clear preference for XV over VX orders. Finally, the constraint PROALIGN, which requires that pronouns precede full NPs, explains why $S_{full}O_{pro}$ is less acceptable than $S_{pro}O_{full}$, while $O_{full}S_{pro}$ is less acceptable than both $O_{full}S_{full}$ and $O_{pro}S_{full}$ (see Figure 4.11).

4.5.5.2. Constraints in Context

The behavior of VERBFINAL was replicated in the context condition. We found that a violation of VERBFINAL leads to serious unacceptability; no context effects were attested for VERBFINAL, which indicates that we are dealing with a hard constraint.

On the other hand, we found an interaction of PROALIGN and context. The prediction that pronouns have to precede full NPs is born out in the all focus context, but in the S focus and O focus contexts, the effect of PROALIGN disappears. This might indicate that PROALIGN is only valid if the context fails to provide an antecedent for the pronoun. According to the context hypothesis developed in Experiment 8 (see also Section 4.1.1), this context effect is an indication that PROALIGN is a soft constraint. Note that the interaction of PROALIGN with context does not readily follow from existing accounts of word order variation in German (Müller 1999; Uszkoreit 1987).

We also provided evidence for the constraint GROUNDALIGN that requires ground NPs to be peripheral. In the S focus context, we found an overall preference for SO, even though SO violates GROUNDALIGN, while OS satisfies it. In the O focus context, however, we found that the SO preference is increased, which can be explained by the fact that SO satisfies GROUNDALIGN in the O focus context, while OS violates it. While this observation provides support for the validity of GROUNDALIGN, the overall SO preference (even if it is disfavored by the context) seems to indicate that the effect of GROUNDALIGN is weak compared to the influence of NOMALIGN, i.e., NOMALIGN should receive a higher ranking than GROUNDALIGN.

4.5.5.3. Constraint Ranking

In the null context condition, we investigated constraint ranking by conducting a series of planned comparisons for single violations of NOMALIGN, VERBFINAL, and PROALIGN in the null context condition. We found that VERBFINAL was ranked higher than both NOMALIGN and PROALIGN. These two constraints, however, did not differ in their ranking. Furthermore, we can assume that NOMALIGN is ranked higher than GROUNDALIGN, based on the fact GROUNDALIGN effects are weak compared to NOMALIGN effects (see above). Hence we arrive at the following overall constraint hierarchy:

(4.30) $\text{VERBFINAL} \gg \{\text{PROALIGN}, \text{NOMALIGN}\} \gg \text{GROUNDALIGN}$

Recall that hard constraints are expected to lead to serious unacceptability, while soft constraints cause only mild unacceptability. This is compatible with the hierarchy in (4.30) if we assume that VERBFINAL is a hard constraint, while PROALIGN, NOMALIGN, and GROUNDALIGN are soft constraints. Also, the context effects support this classification: we found clear evidence that VERBFINAL is context-independent (and thus hard), while the PROALIGN was clearly context-dependent (and thus soft). No context effects were found for

NOMALIGN, while the contextual status of GROUNDALIGN remained unclear. Note however, that context effects are not a *necessary* property of soft constraints.

The ranking in (4.30) is partly compatible with the one proposed by Müller (1999), who stipulates NOM $\gg$ FOC, where his NOM and FOC correspond to our NOMALIGN and (approximately) GROUNDALIGN (see the discussion in Section 3.7.1). However, Müller (1999) classifies his equivalent of PROALIGN as a hard constraint (a grammatical constraint in his terminology). This is not supported by our data, which showed that PROALIGN does not cause categorical unacceptability; we found that PROALIGN is ranked equal to NOMALIGN, which Müller considers a soft constraint (a markedness constraint in his terminology). Another point in case are the context effects that emerged for PROALIGN. According to Müller, context effects are characteristic of markedness (soft) constraints, but not of grammaticality (hard) constraints.

On the other hand, VERBFINAL (not explicitly dealt with by Müller) seems to be a genuine hard constraint. Its violation leads to strong unacceptability in all contexts, independently of which other constraints are violated.

4.5.5.4. Constraint Interaction

In Experiments 4–6 and 8, we found evidence for the fact that constraint violations are cumulative, i.e., that the degree of unacceptability of a stimulus increases with the number of violations incurred. In the present experiment, this finding was confirmed based on a series of planned comparisons on multiple constraint violations in the null context condition. We found clear evidence for the cumulativity of the constraints PROALIGN, NOMALIGN, and VERBFINAL. This is in line with the results from Experiment 6, where the cumulativity of PROALIGN, NOMALIGN, and DATALIGN was demonstrated.

Furthermore, it was shown that lower ranked constraints such as NOMALIGN and PROALIGN can gang up against higher ranked ones such as VERBALIGN: a post-hoc test found that a combined violation of NOMALIGN and PROALIGN is only marginally different from a single violation VERBALIGN. This is consistent with the ganging up effects already demonstrated in Experiments 4 and 5 and constitutes evidence against OT-style strict domination of constraints in a new syntactic domain (word order). It also confirms that soft constraints (NOMALIGN and PROALIGN) can gang up against hard ones (VERBALIGN), in line with the findings of Experiment 5.

4.5.6. Conclusions

The results of the present experiment were fully consistent with the results obtained in Experiment 6. We provided evidence for the word order constraints NOMALIGN and PROALIGN and confirmed that they have the same rank. We also showed that the two additional word or-

der constraints VERBFINAL and GROUNDALIGN have a significant effect on acceptability and established a ranking for these two constraints.

The main aim of the present experiment was to supply evidence for the fact that soft constraints are context-dependent, while hard constraints context-independent, i.e., immune to context effects (see Experiments 7–9). Based on the constraint ranking, we hypothesized that NOMALIGN, PROALIGN, and GROUNDALIGN are soft constraints, while VERBFINAL is a hard constraint. We demonstrated that context effects occur for PROALIGN, but not for VERBFINAL and NOMALIGN. The contextual status of GROUNDALIGN remained unclear. This result is compatible with the hypothesis that only soft constraints are context-dependent. Note however, that it also means that context effects are a sufficient, but not necessary, feature of soft constraints.

In the following two experiments, we will continue our investigation of context effects on word order, providing additional evidence for the hypothesis that context effects are a diagnostic of the soft/hard dichotomy. We will add a crosslinguistic dimension to this investigation by presenting data from Greek, a language that exhibits considerably more word order freedom than German. Furthermore, Experiment 12 is designed to extend the range of data that we consider by dealing with spoken instead of written stimuli, thus allowing us to investigate of the interaction of word order and phonology.

4.6. Experiment 11: Effect of Clitic Doubling, Verb Position, and Context on Word Order

In Experiments 1–3 we observed that the distinction between soft and hard constraint violations manifests itself in crosslinguistic (crossdialectal) effects. These effects were further discussed in Section 4.1.2, where we arrived at the hypothesis that crosslinguistic variation cannot affect the type of a constraint (soft or hard). However, the results of Experiments 1–3 were restricted to a single linguistic phenomenon (unaccusativity and unergativity as manifested in auxiliary selection and impersonal passive formation) and the scope of crosslinguistic variation we considered was limited (two dialects of German).

Experiments 11 and 12 were designed to extend the investigation of crosslinguistic variation and gradience to word order preferences. Word order was already the subject of Experiments 6 and 10, were we dealt with ordering preferences in German, a language with semi-free word order. Experiments 11 and 12 will extend these by investigating an extended set of word order constraints and providing data from Greek, a free word order language. In addition, Experiment 12 will employ spoken stimuli, which will enable us to investigate the effects of accent placement on word order, thus extending the range of data considered.

Apart from providing more evidence for the crosslinguistic behavior of soft and hard constraints, Experiments 11 and 12 will also investigate context effects on word order prefer-

ences, thus accumulating additional evidence for the hypothesis (arrived at in Experiments 7–10) that soft constraint violations are context-dependent, while hard violations are context-independent.

4.6.1. Background

We will require a considerable amount of linguistic background to be able to discuss the results of Experiments 11 and 12. A general overview of Information Packaging will be provided in Section 4.6.1.1. Information Packaging is the framework in which we will attempt to explain the interaction of syntactic and phonological constraints with contextual factors. Following up on this, Section 4.6.1.2 will then present the basic facts about Information Packaging in Greek. In Section 4.6.1.3 will define the set of constraints on which our discussion of Experiments 11 and 12 is based.

We will not attempt to provide a full overview of previous experimental research on Information Structure. Note, however, that previous work as has either focussed on the information structural role of intonation (e.g., Birch and Clifton 1995) or on syntactic markers of Information Structure (such as clefting, e.g., Vion and Colas 1995). Experiments 11 and 12 try to integrate these two strands of research. They constitute the first experimental attempt to clarify the interaction of phonology and syntax in marking Information Structure in a free-word order language such as Greek.

4.6.1.1. Information Structure in English

This section gives an overview of the primitives of the framework of Information Structure introduced by Vallduví (1992). Building on previous work of, among others, Chafe (1976, 1983), and Prince (1986), Vallduví (1992) views a sentence as conveying information that updates the hearer's knowledge-base or *information state*. Each sentence constitutes an *instruction* indicating to the hearer *what* information to add, *where* to add it, and *how*. These instructions are encoded in the *Information Structure* of a sentence, which consists of the following primitives:

(4.31) Sentence := {Focus, Ground}
Ground := {Link, Tail}

Focus conveys the new information of the sentence, whereas *ground* anchors the new information to the hearer's current information state. Ground is further subdivided into *link* and *tail*: link points to the locus of update in the hearer's information state, i.e., to where the new information should be added. Tail indicates *how* information should be added.[7] The three primitives focus, link, and tail combine to yield four instruction types: *all focus*, *link-focus*, *focus-tail*,

[7] Vallduví's tripartite organization of Information Structure combines previous distinctions such as *theme-rheme*, *topic-comment*, and *ground-focus* (Halliday 1967; Reinhart 1982). In particular, his link corresponds to traditional notions of topic or theme.

and *link-focus-tail*. Below we explicate briefly the function of each of these instruction types. Possible realizations of the four instruction in English are exemplified in (4.32); SMALL CAPITALS indicate main sentential stress, **boldface** marks secondary stress. Focus is indicated using square brackets and subscript F.

(4.32) a. **All Focus**
The president has a weakness.
[$_F$ He hates CHOCOLATE].

b. **Link-Focus**
Tell me about the people in the White House. Anything I should know?
The **president** [$_F$ hates CHOCOLATE].

c. **Focus-Tail**
You shouldn't have brought chocolates for the president.
[$_F$ He HATES] chocolate.

d. **Link-Focus-Tail**
And what about the president? How does HE feel about chocolate?
The president [$_F$ HATES] chocolate.

(Vallduví and Engdahl 1996)

All instruction types contain a focus part, since every sentence has an update potential. The presence of ground segments depends on the knowledge shared between the interlocutors in the previous discourse, i.e., on the context in which the sentence is uttered. Let us consider the above examples in more detail. Sentence (4.32a) involves an all focus instruction which updates the information about *the president.* Note that *the president* has already been activated as a locus of update by the context in which (4.32a) appears. Hence (4.32a) does not contain a link.[8] Similarly, the locus of update is inherited from the previous context in (4.32c), which also conveys a "link-less" instruction. On the other hand, example (4.32b) specifies *the president* as its link, i.e., the locus of update. (4.32b) instructs the hearer to add the new condition *hates chocolate* to this locus. Finally, example (4.32d) also specifies *the president* as the locus of update, but conveys a different update instruction (tail). It instructs the hearer to search for a condition of the form *likes chocolate* and replace the predicate *likes* with *hates*. Note that the same instruction is also encoded by the tail in (4.32c).

Let us turn to the linguistic means by which English marks the different components of Information Structure. All of the examples in (4.32) involve the same word order. However, they differ in their intonational pattern: English relies on prosodic means for encoding Information Structure. Focused segments are associated with the main sentential stress (A accent). This is true of *narrow focus* as in (4.32d), but also *broad focus* as in (4.32b). (Narrow and broad focus are descriptive terms: narrow focus denotes NP or verb focus, while broad focus refers to

[8] Pronouns do not contribute to the Information Structure of a sentence. They are just syntactic placeholders (Vallduví 1995).

VP or S focus). Any accented constituent can be interpreted as narrow focus, while only accent on the rightmost complement can give rise to a broad focus interpretation (Ladd 1996; Vallduví 1992).[9] English provides intonational marking not only for foci, but also for links. Links like *the president* in (4.32b) receive secondary stress. (This accent is referred to as B accent in the theoretical literature, see Ladd 1996; Pierrehumbert and Hirschberg 1990; Steedman 1991. For further discussion on the realization of Information Structure in English see also Bolinger 1978, 1989; Rochemont 1986; Selkirk 1984.)

Note that our choice of the framework of Information Structure does not bear directly on the claims we make in subsequent sections. It mainly serves as a theoretical background against which we discuss the phenomena at hand. However, Vallduví's assumption that Information Structure forms an independent grammatical level, interacting with both syntax and phonology, is compatible with the model of constraint interaction we will advocate in the remainder of this chapter. Note further that the experiments we will report only investigate the ground-focus distinction and do not explore the distinction between link and tail. This restriction was necessary to keep the experimental design to a manageable size. For completeness, the present section introduced all three Information Structure primitives.

4.6.1.2. Information Structure in Greek

As in English, accent placement plays a central role in the realization of Information Structure in Greek. However, in addition to phonological resources, Greek also employs syntactic resources such as word order and clitic doubling to realize Information Structure. In this section, we briefly present how phonological and syntactic devices combine to yield various Information Structure instructions in Greek.

We will employ a notation that uses capitalization to indicate accent, e.g., svO indicates the order subject-verb-object with accent on the object. The same order with clitic doubling is denoted as sclvO. We use an all capital notation where we disregard accent, such as in SVO.

Let us first consider cases of narrow focus as in (4.33) where the subject NP *o Yanis* is focused. Focused NPs are accented, as in English, but, unlike English, their order is not fixed; they may appear either preverbally (see (4.33b)) or postverbally (see (4.33a)). (According to the literature, preverbal focus is more likely to be associated with a contrastive reading and therefore might be considered slightly more marked than postverbal focus, see Alexopoulou 1998; Tsimpli 1995.) It is not only focused NPs that are free to appear either preverbally or postverbally; so do ground NPs (*ti Maria* in (4.33)). The standard literature assumes that preverbal ground NPs realize Vallduvían links (or topics in the traditional sense), while postverbal ones are interpreted as tails (Alexopoulou 1998; Anagnostopoulou 1994; Philippaki-Warburton 1985; Schneider-Zioga 1994; Tsimpli 1995; Valiouli 1994). Hence example (4.33a) realizes a

[9] Technically, it is the most oblique rather than the rightmost NP in English (Vallduví and Engdahl 1996). However, the most oblique NPs in English are typically the rightmost ones.

link-focus instruction in Vallduví's terms, while (4.33b) realizes a *focus-tail* instruction.

(4.33) **Subject Focus**
Pios apelise ti Maria?
"Who fired Maria?"
a. Ti Maria tin apelise [$_F$ o YANIS].
the Maria-ACC her-CL fired-3SG [$_F$ the Yanis-NOM]
"Yanis fired Maria."
b. [$_F$ O YANIS] (tin) apelise ti Maria.
[$_F$ the Yanis-NOM] her-CL fired-3SG the Maria-ACC

Both ground and focused NPs can alternate between preverbal and postverbal positions. Ground NPs, however, tend to appear in peripheral positions while focused ones are preferred adjacent to the verb (Alexopoulou 1998; Schneider-Zioga 1994; Tsimpli 1995). Thus, the clvSo order in (4.34b), where the focused subject NP is adjacent to the verb, is a felicitous answer to (4.33); in contrast, the clvoS sentence in (4.34a), where the focused NP is dislocated to the right periphery of the clause, is infelicitous.

(4.34) **Subject Focus**
Pios apelise ti Maria?
"Who fired Maria?"
a. ?Tin apelise ti Maria [$_F$ o YANIS].
her-CL fired-3SG the Maria-ACC [$_F$ the Yanis-NOM]
"Yanis fired Maria."
b. Tin apelise [$_F$ o YANIS] ti Maria.
her-CL fired-3SG [$_F$ the Yanis-NOM] the Maria-ACC

Similarly, preverbal ground NPs typically precede preverbal focus (see (4.35a)). (In fact, various authors consider examples like (4.35b), where the link follows the focused NP, ungrammatical, see Tsimpli 1995; Tsiplakou 1998.)

(4.35) a. Ti Maria [$_F$ o YANIS] tin apelise horis kamia proidopiisi.
the Maria-ACC [$_F$ the Yanis-NOM] her-CL fired-3SG without any warning
"Yanis fired Maria without any warning."
b. ?[$_F$ O YANIS] ti Maria tin apelise horis kamia proidopiisi.
[$_F$ the Yanis-NOM] the Maria-ACC her-CL fired-3SG without any warning

Note finally, that the ground object NP in (4.33) is marked by an additional clitic pronoun, attached to the verb. We will use the term *clitic doubling* to refer to a configuration where the object NP is co-indexed with such a clitic, irrespective of the position of the object NP.[10] When the NP is postverbal, doubling is optional (as indicated by the brackets around the clitic *tin*

[10] Our term subsumes clitic doubling and clitic left dislocation which are used for postverbal and left dislocated clitic doubled NPs, respectively.

in (4.33b)); but doubling tends to be obligatory when the NP is dislocated to the left (we return to this issue later in this section).

The ground-focus partition is thus realized in Greek through the exploitation of diverse structural resources: accent placement (on the focused constituent), word order (focused NPs are adjacent to the verb, ground ones are dislocated to peripheral positions), and clitic doubling (object ground NPs are preferred doubled). The interaction between these structural devices follows a consistent pattern, independent of the grammatical function of the focused or ground NP (modulo the fact that clitic doubling is only available for objects in Greek). Consider the examples in (4.36) which demonstrate a narrow focus reading for the object NP:

(4.36) **Object Focus**

Pion apelise i Maria?

"Who did Maria fire?"

a. I Maria apelise [$_F$ to YANI].
the Maria-NOM fired-3SG [$_F$ the Yanis-ACC]
"Maria fired Yanis."

b. [$_F$ To YANI] apelise i Maria.
[$_F$ the Yanis-ACC] fired-3SG the Maria-NOM

c. *[$_F$ To YANI] ton apelise i Maria.
[$_F$ the Yanis-ACC] him-CL fired-3SG the Maria-NOM

Again, the accent falls on the focused NP *to Yani*, which can appear either preverbally (see (4.36b)) or postverbally (see (4.36a)), while the ground subject NP *i Maria* is unaccented. Again, the focused object NP is preferred adjacent to the verb, as indicated by the felicity of the vOs order in (4.37a) compared to the vsO order in (4.37b), as answers to the question in (4.37).

(4.37) **Object Focus**

Pion apelise i Maria?

"Who did Maria fire?"

a. Apelise [$_F$ to YANI] i Maria.
fired-3SG [$_F$ the Yanis-ACC] the Maria-NOM
"Maria fired Yanis."

b. ?Apelise i Maria [$_F$ to YANI].
fired-3SG the Maria-NOM [$_F$ the Yanis-ACC]

Focused objects cannot be doubled, as the unacceptability of (4.36c) indicates. In fact, sentences with accent on a clitic doubled object are unacceptable, irrespective of the context they appear in (Agouraki 1993; Alexopoulou 1998), as illustrated by (4.38):

(4.38) a. *To YANI ton ida.
the Yani-ACC him-CL saw-1SG
"I saw Yanis."

b. *Ton ida to YANI.
him-CL saw-1SG the Yani-ACC

The general unacceptability of accented doubled objects is due to the conflicting information structural requirements imposed on these objects. Accent marks the object as focus, while doubling marks it as ground.

Verb focus is marked with accent on the verb as in (4.39). Again, the ground NPs can appear either preverbally (see (4.39a)) or postverbally (see (4.39b)), while the object NP is preferred doubled:

(4.39) **Verb Focus**
Ti ekane o Yanis me to aftokinito?
"What did Yanis do with the car?"
a. O Yanis [$_F$ to PULISE] to aftokinito.
the Yanis-NOM [$_F$ it-CL sold-3SG] the car-ACC
"Yanis sold the car."
b. To aftokinito [$_F$ to PULISE] o Yanis.
the car-ACC [$_F$ it-CL sold-3SG] the Yanis-NOM

Let us now turn to all focus instructions. VSO is standardly considered the most natural response to an all focus question (see (4.40)). In fact, the naturalness of VSO in this context has been part of the argument for the standard analysis of VSO as the basic order of Greek (Agouraki 1993; Alexopoulou 1998; Philippaki-Warburton 1985; Tsimpli 1995). In such contexts, accent falls on the rightmost constituent. In this respect, Greek seems to pattern with English; in both languages an all focus interpretation arises from accent on the rightmost NP.

(4.40) **All Focus**
Kana neo?
"Any news?"
[$_F$ pulise o Yanis to AFTOKINITO].
[$_F$ sold-3SG the Yanis-NOM the car-ACC]
"Yanis sold the car."

It is worth mentioning here that, as Vallduví and Engdahl (1996) note, questions introducing an all focus context (*What happened/Any news?)* can also give rise to VP focus, with the subject or object dislocated to the left periphery of the clause. Indeed, svO (see (4.41a)) and oclvS (see (4.41b)), instantiating a link/topic-focus Information Structure, are also felicitous answers to an all focus question.

(4.41) **All Focus**
Kana neo?
"Any news?"

a. O Yanis pulise to AFTOKINITO.
the Yanis-NOM sold-3SG the car-ACC
"Yanis sold the car."

b. Tis fises tha tis stilume AVRIO.
the posters-ACC FUT them-CL send-1PL tomorrow
"We will send the posters tomorrow."

Broad and narrow focus contexts differ significantly in the range of utterances they can accommodate. A broad focus context allows the accommodation of a wider range of ground-focus partitions, while a narrow focus context only accepts sentences with a ground-focus partition strictly corresponding to the expectations it imposes. For instance, example (4.42) can be a felicitous answer to a question like (4.40), even though it does not directly correspond to an all focus instruction; rather than the rightmost constituent, the accent falls on the verb, while the object NP is doubled. This sentence is acceptable in a context where the two interlocutors share the knowledge that Yanis was expected to sell his car. However, even if such knowledge is shared by the two speakers, such a sentence would not be acceptable as an answer to an object focus question like *What did Yanis sell?*; only a ground (object) focus instruction constitutes a felicitous answer for this question.

(4.42) To PULISE to aftokinito o Yanis.
it-CL sold-3SG the car-ACC the Yanis-NOM
"Yanis sold the car."

It is worth mentioning here that the wider range of answers satisfying an all focus question yields higher freedom in the linguistic realization of these answers. Thus, most orders (SVO, OVS, and the verb initial orders) are acceptable, while the accent may shift from the rightmost clause boundary to the left.

To summarize, Information Structure in Greek is realized through a combination of phonological and syntactic means, captured by the following descriptive generalizations:

(4.43) **Descriptive Generalizations on Information Structure**

a. **Phonology:** (i) Accented constituents are (part of) focus; ground elements bear no accent; (ii) accent on the rightmost NP gives rise to a broad focus interpretation.

b. **Word Order:** ground constituents are peripheral.

c. **Clitic Doubling:** doubled objects are ground.

Note that, while it is true that ground NPs are peripheral, it is not always the case that focused NPs are adjacent to the verb. Adjacency is observed in cases of narrow focus but not always in all focus instructions. In these instructions accent falls on the rightmost NP which, very often, is not adjacent to the verb (e.g., vsO or voS).

In the following, we will briefly comment on two more restrictions on word order in Greek. First, as mentioned earlier, preverbal ground objects (see (4.33a)) should be doubled,

while doubling is optional with postverbal ground NPs. The obligatoriness of the clitic in examples like (4.44) has been a matter of controversy in the literature. Examples like these are often judged less acceptable when they lack doubling, with some authors judging them unacceptable (Tsiplakou 1998). One goal of our experimental study is to settle the data disputes that surround these examples.

(4.44) a. To aftokinito ?(to) pulise o YANIS.
the car-ACC it-CL sold-3SG the Yanis-NOM
"Yanis sold the car."

b. Tin Maria ?(tin) apelise o YANIS.
the Maria-ACC her-CL fired-3SG the Yanis-NOM
"Yanis fired Maria."

In Section 4.6.1.3 we will postulate a constraint requiring preverbal objects to be doubled and we will provide evidence supporting this constraint in Section 12.

The second issue concerns verb final word orders. So far we have only considered orders in which the verb appears in either initial or medial position. Verb final orders (SOV and OSV), while grammatical, are generally perceived as less acceptable:

(4.45) a. ?Tis Marias o Yanis / o Yanis tis Marias tis milise.
the Maria-GEN the Yanis-NOM / the Yanis-NOM the Maria-GEN her-CL talked-3SG
"Yanis talked to Maria."

b. ?Ta pedia o Yanis / o Yanis ta pedia ta ide.
the kids-ACC the Yanis-NOM / the Yanis-NOM the kids-ACC them-CL saw-3SG
"Yanis saw the kids."

However, SOV and OSV improve to full acceptability if more material is added after the verb:

(4.46) a. Tis Marias o Yanis / o Yanis tis Marias tis ipe oti de
the Maria-GEN the Yanis-NOM / the Yanis-NOM the Maria-GEN her-CL said-3SG that not
theli na pai stin Ameriki.
want SUBJ go to-the America
"Yanis told Maria that he doesn't want to go to America."

b. Ta pedia o Yanis / o Yanis ta pedia ta vlepi mono
the kids-ACC the Yanis-NOM / the Yanis-NOM the kids-ACC them-CL see-3SG only
otan den ehi dulia to Savatokiriako.
when not have work the weekend
"Yanis sees the kids only when he doesn't have work during the weekend."

To account for the reduced acceptability of (4.45) compared with (4.46), we will assume a constraint that penalizes verbs that occur at clause final positions (see Section 4.6.1.3).

To summarize: all of the factors discussed in this section are expected to have a significant effect on the acceptability of a given word order. This includes the information structural

factors listed in (4.43), as well as the restrictions on preverbal objects and on verb final sentences stated above. However, it quickly becomes evident that not all of these factors play an equally important role. One of the main goals of Experiments 11 and 12 is to identify the nature of the interaction between these factors and to quantify the effect of each of them. Before we present the experimental results, some preliminary observations are in order.

Accent placement appears as the most important factor in the realization of the ground-focus partition as it is both obligatorily and unambiguously associated with (at least part of) focus. Word order, on the other hand, appears as a comparatively weak factor. In the absence of accent and clitic doubling, a given order may give little or no indication of the ground-focus partition; for example, SVO and OVS can realize a link-focus or focus-ground partition, depending on accent placement and doubling (svO/oclvS and Svclo/Ovs). Similarly, VSO can realize an all focus sentence or allow a narrow focus interpretation for the subject NP (vsO and vSclo respectively). Unlike word order, clitic doubling is unambiguously associated with a ground interpretation of objects. However, unlike accent, doubling is not necessary for the realization of ground NPs, and its effect is restricted to objects.

In Section 4.6.1.3, we will introduce a set of grammatical constraints based on the generalizations presented above. We expect that the experimental results will show that all these factors play a role in the realization of Information Structure, while the magnitude of the effect on acceptability judgments caused by each of these factors will reflect its relative importance. More precisely, we expect violations of accent placement to induce the strongest effect. Given its unambiguous association with focus, accent placement provides hearers with a strong cue for the Information Structure of a sentence. The restriction that doubled NPs cannot function as foci is also expected to produce strong effects. Just as accent placement, clitic doubling is an unambiguous marker of Information Structure. Violations of word order preferences, on the other hand, are expected to trigger weak effects; given its ambiguity, word order is an additional, but rather unreliable cue for detecting the ground-focus partition of a sentence. Note as well that, due to the ambiguity of word order, some word orders will satisfy the information structural requirements of several contexts.

4.6.1.3. Constraints on Information Structure

Based on the observations outlined in Section 4.6.1.2, we propose a set of linguistic constraints that govern the realization of Information Structure in Greek. The purpose of these constraints is to facilitate a systematic discussion of the data and to exemplify how a constraint-based approach can capture basic aspects of the experimental results. We will restrict ourselves to a fairly descriptive formulation of the constraints (for more linguistically sophisticated accounts of Information Structure and word order in an optimality theoretic setting, see Choi 1996; Müller 1999; Samek-Lodovici 1996).

The constraints in (4.47) are based on our generalizations on Information Structure

summarized in (4.43), and on the observations regarding clitic doubling and verb final orders discussed at the end of Section 4.6.1.2.

(4.47) **Constraints on Word Order and Information Structure**

a. GROUNDALIGN (GAGN): ground constituents have to be peripheral.
b. DOUBLEGROUND (DOUG): clitic doubled objects have to be interpreted as ground.
c. ACCENTALIGN (ACCAGN): accent has to fall on the rightmost constituent.
d. ACCENTFOCUS (ACCF): accented constituents have to be interpreted as focus.
e. DOUBLEALIGN (DOUAGN): preverbal objects have to be clitic doubled.
f. VERBALIGN (VAGN): the verb must not be right peripheral.

The first two constraints impose restrictions on the syntactic/morphological realization of Information Structure. GROUNDALIGN encodes the restriction that ground NPs should appear either to the left or right periphery of the clause. We use the term "periphery" descriptively, to refer to clause initial and clause final NPs. Note that this restriction is not biconditional; peripheral NPs do not necessarily belong to the ground part of the sentence. Furthermore, the association of doubled NPs with a ground interpretation is captured by DOUBLEGROUND.

While GROUNDALIGN and DOUBLEGROUND encode syntactic/morphological restrictions on ground elements, ACCENTFOCUS and ACCENTALIGN are phonological constraints on the realization of focused NPs. ACCENTFOCUS associates an accented constituent with a focus interpretation. It applies to all Information Structures, i.e., both in narrow and broad focus contexts. Moreover, ACCENTFOCUS is insensitive to other structural properties of the relevant constituent (e.g., whether the constituent is an NP or not, whether it appears preverbally or postverbally). ACCENTALIGN, on the other hand associates accent placement with clause structure (the right clause boundary).

The first four constraints restrict the realization of Information Structure (see (4.43)), while the last two constraints impose restrictions on word order, independent of information structural factors. DOUBLEALIGN requires preverbal objects to be doubled, while VERBALIGN penalizes verb final orders. (For a more detailed motivation of these two constraints, see Section 4.6.1.2.)

4.6.2. Introduction

Experiment 11 has a double purpose. Firstly, it investigates the basic claim that word order plays an information structural role in a free word order language like Greek. We elicit acceptability judgments for a variety of word orders and contexts, which allows us to examine the interaction of word order and context. Secondly, the experiment is designed to assess the effect of three constraints: the word order constraint VERBALIGN, the constraint on clitic doubling DOUBLEALIGN, and the constraint GROUNDALIGN regulating the interaction of word

order and Information Structure (see (4.47) for details). The experiment includes sentences that violate one or more of these constraints, and the prediction is that such violations lead to a reduction in acceptability.

The experimental design includes two factors: word order (*Ord*) and context (*Con*). Six word orders were tested: SVO, OVS, VSO, VOS, SOV, and OSV, as illustrated by the following examples:

(4.48) a. **SVO:** O Tasos tha diavasi tin efimerida.
the Tasos-NOM will read-3SG the newspaper-ACC
"Tasos will read the newspaper."
b. **OVS:** Tin efimerida tha diavasi o Tasos.
c. **VSO:** Tha diavasi o Tasos tin efimerida.
d. **VOS:** Tha diavasi tin efimerida o Tasos.
e. **SOV:** O Tasos tin efimerida tha diavasi.
f. **OSV:** Tin efimerida o Tasos tha diavasi.

Clitic doubled sentences were not included in this experiment, in order to keep the design at a manageable size. Note that DOUBLEALIGN can be tested on structures that do not contain doubling: for instance, OVS (that violates DOUBLEALIGN) can be compared with SVO (that does not violate DOUBLEALIGN). (Clitic doubled stimuli were included Experiment 12, allowing a direct comparison of OVS with OclVS.)

For the context factor we employed a question context to establish a pattern of ground and focus information, a technique that is widely used in the theoretical literature (e.g., Vallduví 1992). A total of five contexts were used: null, all focus, subject focus, object focus, and verb focus. As an example, consider the contexts for the sentences in (4.48):

(4.49) a. **Null**
b. **All Focus:** Ti tha gini?
"What will happen?
c. **S Focus:** Pios tha diavasi tin efimerida?
"Who will read the newspaper?"
d. **O Focus:** Ti tha diavasi o Tasos?
"What will Tasos read?"
e. **V Focus:** Ti tha kani o Tasos me tin efimerida?
"What will Tasos do with the newspaper?"

The null context was included as a control condition, allowing us to study how subjects react in the absence of any contextual information.

4.6.3. Predictions

4.6.3.1. Constraints out of Context

The general prediction is that some word orders are more acceptable than others. Hence we expect to find a main effect of *Ord* (word order).

Furthermore, the constraints in (4.47) allow us to make detailed predictions about the acceptability of individual orders. If a given structure violates one of the constraints in (4.47), then we predict its acceptability to be reduced compared a structure that does not incur this constraint violation. Only the constraints VERBALIGN, DOUBLEALIGN, and GROUNDALIGN are relevant for the present experiment. The other three constraints (DOUBLEGROUND, ACCENTALIGN, and ACCENTFOCUS) deal with clitic doubling and accent placement, and will be investigated in Experiment 12.

VERBALIGN requires that verbs must not occur at the right periphery of a sentence (i.e., sentence initially or sentence finally). This constraint is violated by verb final sentences, where the verb appears clause finally (SOV and OSV sentences in our stimulus set). Hence we expect these orders to be reduced in acceptability. The constraint DOUBLEALIGN requires preverbal objects to be clitic doubled. This constraint is violated by OVS, SOV, and OSV. These orders contain preverbal objects that are not doubled and hence are predicted to be reduced in acceptability. The constraint GROUNDALIGN requires ground constituents to be sentence peripheral. This constraint does not apply in the null context condition, where no information about ground and focus is available.

4.6.3.2. Constraints in Context

The general prediction in the context condition is that context has an influence on word order preferences. Hence we expect an interaction of *Ord* (word order) and *Con* (context).

Again, the constraints in (4.47) make predictions based on individual constraint violations. On the one hand, we expect effects from VERBALIGN and DOUBLEALIGN. These are syntactic constraints that are not subject to information structural effects. Hence their effects in the context condition should be the same as in the null context condition, i.e., VERBALIGN should disfavor verb final orders (SOV and OSV), while DOUBLEALIGN should disfavor OVS, SOV, and OSV, as these orders include preverbal non-doubled objects.

As for the interaction of word order and context, we expect that the order preferences for each context will reflect the optimal realization of the Information Structure required for this context. More specifically, the constraint GROUNDALIGN predicts that orders with non-peripheral ground constituents will be reduced in acceptability. In the following, we will discuss the predictions for each context.

All Focus Context There are no ground constituents in the all focus context, hence GROUNDALIGN is vacuously satisfied. The order preferences only depend on VERBALIGN and DOUBLEALIGN. The all focus context is therefore predicted to exhibit the same pattern of word order preferences as the null context.

S Focus Context In the S focus context, the subject is in focus, while the object is part of ground. VOS violates GROUNDALIGN, as the object is non-peripheral, and is thus predicted to be less acceptable than SVO, OVS, and VSO, which satisfy GROUNDALIGN. GROUNDALIGN is also violated in SOV, which is therefore predicted to be less acceptable than OSV (both orders also violate VERBALIGN, and hence should be generally low in acceptability).

O Focus Context In the O focus context, the object is in focus, while the subject is part of ground. This means that GROUNDALIGN is violated in VSO, where the subject is non-peripheral. Hence VSO should be dispreferred compared to SVO, OVS, and VOS, which satisfy GROUNDALIGN. SOV also incurs a GROUNDALIGN violation, and hence should be less acceptable than OSV (both orders also violate VERBALIGN).

V Focus Context In the V focus context, the verb is in focus, while both the subject and the object are ground constituents. According to GROUNDALIGN, both NPs have to appear in peripheral positions, i.e., clause final or clause initial. It follows that all orders except SVO and OVS violate GROUNDALIGN. Thus, VSO, VOS, SOV and OSV are predicted to be reduced in acceptability compared with SVO and OVS. However, as OVS violates DOUBLEALIGN, SVO is expected to be the best order. The two final orders, SOV and OSV should be least acceptable: unlike VSO, VOS, and OVS, they violate three constraints (VERBALIGN, DOUBLEALIGN and GROUNDALIGN).

4.6.3.3. Constraint Types

The present experiment also allows us to determine the type of the three constraints under investigation. We can diagnose whether a constraint is hard or soft based on three criteria: constraint strength, context effects, and crosslinguistic effects.

Experiment 10 dealt with the effects of the constraint GROUNDALIGN in German and demonstrated that GROUNDALIGN is a soft constraint. Under the hypothesis that crosslinguistic variation cannot affect the type of a constraint (see Section 4.1.2), this means that GROUNDALIGN is expected to be a soft constraint also in Greek. We therefore expect GROUNDALIGN to be context-dependent and induce only weak acceptability differences. The status of DOUBLEALIGN and VERBALIGN remains to be determined; depending on constraint strength and contextual behavior, these constraints will be classified as either a soft or hard.

In previous experiments, we determined constraint ranking by comparing structures that incur only single constraint violations. Due to the design of the present experiment, this

approach is not possible here; for some constraints, no single violations were included in the stimulus set (an example is the constraint VERBALIGN). The general problem is that the set of constraints we investigate cannot be mapped straightforwardly onto the set of factors used in the experimental design (as was the case in previous experiments). To establish constraint rankings, we will therefore rely on optimality-theoretic *ranking arguments*. A ranking procedure based on ranking arguments will be defined in Chapter 6 and applied to the data from the present experiment in Chapter 7.

4.6.4. Method

4.6.4.1. Subjects

Forty native speakers of Greek participated in the experiment. The subjects were recruited over the Internet by postings to relevant newsgroups and mailing lists. Participation was voluntary and unpaid. Subjects had to be linguistically naive, i.e., neither linguists nor students of linguistics were allowed to participate.

The data of three subjects were excluded because they were bilingual (by self-assessment). The data of one further subject were excluded as she was a speaker of Cypriot Greek.[11] The data of two subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 34 subjects for analysis. Of these, 19 subjects were male, 15 female; five subjects were left-handed, 29 right-handed. The age of the subjects ranged from 21 to 42 years, the mean was 26.7 years.

4.6.4.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials For the experimental items, a full factorial design was used with word order (*Ord*) and context (*Con*) as the two factors (see (4.48) and (4.49) for example stimuli). This yielded a total of $Ord \times Con = 6 \times 5 = 30$ cells. Eight lexicalizations per cell were used, which resulted in a total of 240 stimuli.

A set of 24 fillers was used, designed to cover the whole acceptability range. Six items of each of the following four groups were used: no violation, case violation, phrase structure violation, and agreement violation. The fillers covered a range of word orders, including ones that were not used in the experimental items (e.g., by using null subjects). The contexts for the fillers included *wh*-questions (both adjunct and complement questions) and *yes-no*-questions.

[11] Cypriot Greek is a dialect that differs considerably from standard Greek. It is not clear whether the differences between Cypriot and standard Greek would affect the current study, but for methodological reasons, it was decided to exclude speakers of Cypriot Greek.

As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

No frequency matching was conducted for the materials in this experiment, as no adequate corpus was available for Greek.

4.6.4.3. Procedure

The method used was magnitude estimation of linguistic acceptability, with the same experimental protocol as in Experiment 1.

Instructions We used a Greek version of the instructions in Experiment 1. Where contextualized stimuli were presented, subjects were told that each sentence would be presented in context, defined as a single sentence preceding the target sentence. Subjects were instructed to judge the acceptability of the target sentence, and to take the context into account in their judgments. The task was illustrated by examples.

Demographic Questionnaire, Training and Practice Phase These were designed in the same way as in Experiment 1.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

A between-subjects design was used to administer the experimental stimuli: subjects in Group A judged non-contextualized stimuli, while subjects in Group B judged contextualized stimuli.

For Group A, two test sets were used: each set contained four lexicalizations for each of the six levels of factor *Ord*, i.e., a total of 24 items. For Group B, eight test sets were used: each set contained one lexicalization for each of the six orders in each of the four contexts, i.e., a total of 24 items. Lexicalizations were assigned to test sets using Latin squares. Two separate Latin squares were applied: one for the null context condition and one for the context condition.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 48 test items: 24 experimental items and 24 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to a group and a test set; 17 subjects were assigned to each group. Instructions, examples, training items, and fillers were adapted for Group B to take context into account.

4.6.5. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

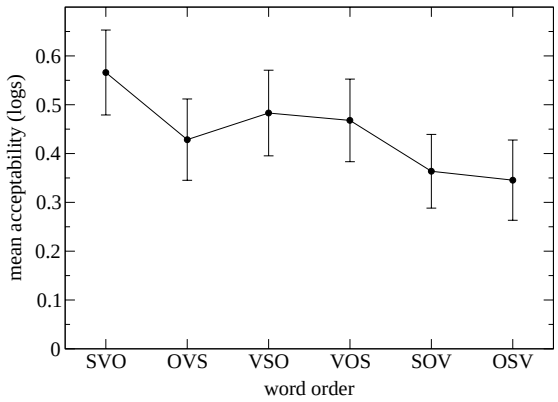

Figure 4.15: Mean judgments for each word order in the null context (Experiment 11)

4.6.5.1. Constraints out of Context

The mean judgments for the null context condition are graphed in Figure 4.15. An ANOVA revealed a significant main effect of word order ($F_1(5,80) = 20.005$, $p < .0005$; $F_2(5,35) = 3.181$, $p = .018$). This confirms our general prediction that some word orders are more acceptable than others, even in absence of context.

A post-hoc Tukey test was carried out for the main effect of *Ord*. This test determines which word orders differ in acceptability and thus allows us to assess the influence of the constraints VERBALIGN and DOUBLEALIGN.

VERBALIGN requires that verbs must not occur clause finally, thus predicting reduced acceptability for the verb final orders SOV and OSV. This was confirmed by the Tukey test, which showed that SOV was significantly less acceptable than SVO (by subjects, $\alpha < .01$, and by items, $\alpha < .05$), VSO (by subjects only, $\alpha < .01$), and VOS (by subjects only, $\alpha < .01$). OSV was significantly less acceptable than SVO (by subjects, $\alpha < .01$, and by items, $\alpha < .05$), VSO (by subjects only, $\alpha < .01$), and VOS (by subjects only, $\alpha < .01$).

The constraint DOUBLEALIGN requires preverbal objects to be clitic doubled, which means that OVS should be reduced in acceptability compared to SVO. This prediction was confirmed by the Tukey test, which showed that OVS was less acceptable than SVO (by subjects only, $\alpha < .01$). Furthermore, we found that OSV was less acceptable than OVS (by subjects only, $\alpha < .05$). Both orders violate DOUBLEALIGN, but OSV is verb final and hence also violates VERBALIGN, which explains the difference in acceptability.

In addition, we found that SVO was more acceptable than the verb initial orders VSO (by subjects only, $\alpha < .01$) and VOS (by subjects only, $\alpha < .01$). This is unexpected, as neither of these three orders violates any constraints, and we would expect them to be equally acceptable. All other differences failed to reach significance.

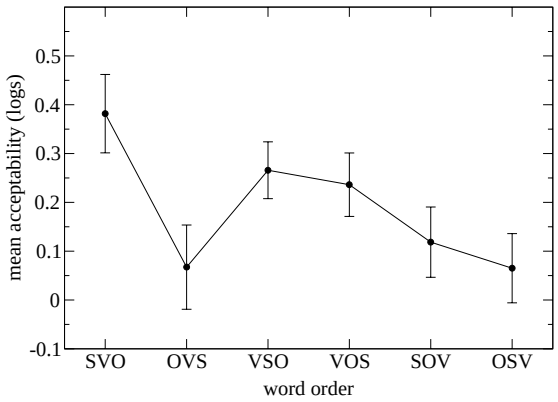

Figure 4.16: Mean judgments for each word order in the all focus context (Experiment 11)

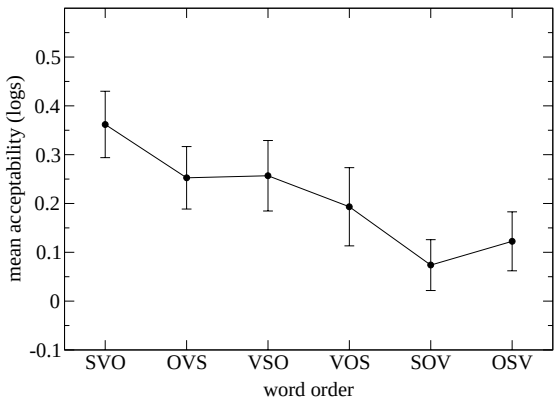

Figure 4.17: Mean judgments for each word order in the S focus context (Experiment 11)

4.6.5.2. Constraints in Context

The mean judgments for the context condition are graphed in Figures 4.16–4.19. As in the null context condition, an ANOVA revealed a significant main effect of word order ($F_1(5,80) = 24.970$, $p < .0005$; $F_2(5,35) = 11.148$, $p < .0005$). A marginally significant main effect of context was also found ($F_1(3,48) = 2.579$, $p = .064$; $F_2(3,21) = 3.275$, $p = .041$). The interaction of word order and context was also significant ($F_1(15,240) = 2.465$, $p = .002$; $F_2(15,105) = 1.969$, $p = .024$), which confirms our general prediction that context has an influence on word order preferences.

A post-hoc Tukey test was carried out on the *Ord* effect to determine the effects of the context independent constraints VERBALIGN and DOUBLEALIGN. The resulting pattern closely matched the one found in the null context condition. Verb final orders were reduced in acceptability, in line with the predictions of VERBALIGN. SOV was significantly less acceptable than SVO ($\alpha < .01$), VSO (by subjects only, $\alpha < .05$), and VOS (by subjects only, $\alpha < .01$).

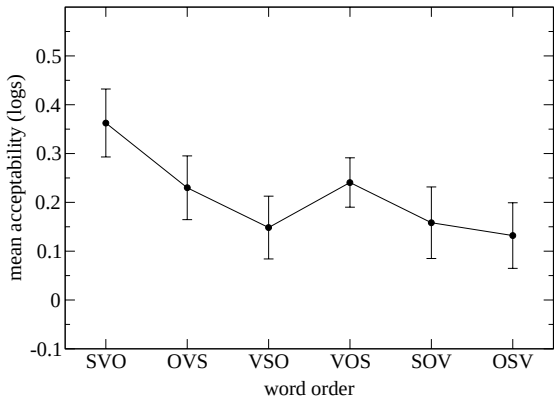

Figure 4.18: Mean judgments for each word order in the O focus context (Experiment 11)

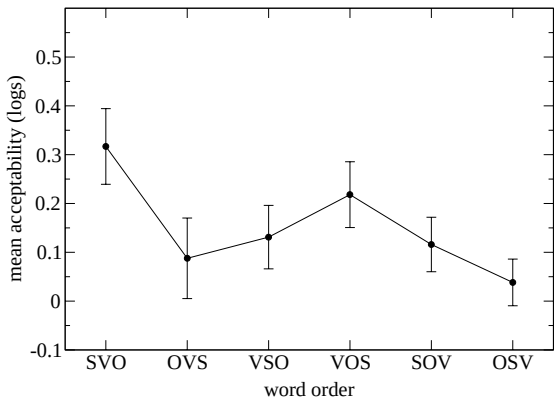

Figure 4.19: Mean judgments for each word order in the V focus context (Experiment 11)

Furthermore, OSV was significantly less acceptable than SVO ($\alpha < .01$), VSO (by subjects only, $\alpha < .01$), and VOS (by subjects, $\alpha < .01$, and by items, $\alpha < .05$). We also found that OVS was less acceptable than SVO ($\alpha < .01$), in line with the predictions of DOUBLEALIGN.

As in the null context condition, SVO was more acceptable than VSO ($\alpha < .01$) and VOS (by subjects, $\alpha < .01$, and by items, $\alpha < .05$). There were no other significant differences.

A further Tukey test was carried for the *Ord*/*Con* interaction to assess the effect of the constraint GROUNDALIGN, which predicts that orders with non-peripheral ground constituents will be reduced in acceptability. We will discuss each context separately.

All Focus Context GROUNDALIGN is vacuously satisfied in an all focus context. Hence we predicted that the all focus context will show the same pattern of order preferences as the null context. This prediction was borne out, as a comparison of Figure 4.15 (null context) and Figure 4.16 (all context) shows.

S Focus Context Here we predicted that VOS, which violates GROUNDALIGN, should be reduced in acceptability compared to SVO, OVS, and VSO, which all satisfy GROUNDALIGN. The Tukey test (see Figure 4.17) provided a partial confirmation: VOS was significantly less acceptable than SVO (by subjects only, $\alpha < .05$). However, the differences between VOS and OVS and between VOS and VSO failed to reach significance.

We also predicted OSV to be preferred over SOV, which violates GROUNDALIGN. Again, this preferences was to small to reach significance. On the other hand, SOV was significantly less acceptable than SVO ($\alpha < .01$), OVS (by subjects only, $\alpha < .05$), and VSO (by subjects only, $\alpha < .05$). OSV was less acceptable than SVO ($\alpha < .01$). These differences are readily explained by the constraint VERBALIGN, which is violated in verb final orders, but not in verb initial and verb medial ones. All other differences were not significant.

O Focus Context Here we predicted VSO (violating GROUNDALIGN) to be less acceptable than SVO, OVS, and VOS (all satisfying GROUNDALIGN). This was partially borne out by the Tukey test (see Figure 4.18) which demonstrated that VSO was significantly less acceptable than SVO (by subjects, $\alpha < .01$, and by items, $\alpha < .05$). However, we failed to find significant differences between VSO and OVS and between VSO and VOS.

We also predicted SOV to be preferred over OSV, which violates GROUNDALIGN. Again, this preference was too small to reach significance. On the other hand, we found that the preference SVO $>$ SOV was significant (by subjects, $\alpha < .01$, and by items, $\alpha < .05$), as well as the preference SVO $>$ OSV ($\alpha < .01$). This is explained by the fact that the verb final orders violate VERBALIGN. There were no other significant differences.

V Focus Context In the V focus context, VSO and VOS violate GROUNDALIGN and hence are predicted to be reduced in acceptability compared to SVO, which satisfies GROUNDALIGN. The Tukey test (see Figure 4.19) confirmed this by showing that SVO was significantly more acceptable than SVO ($\alpha < .05$). The difference between SVO and VOS, however, failed to reach significance. On the other hand, OVS was less acceptable than SVO, readily explained by the fact that OVS violates DOUBLEALIGN. Also, the preference SVO $>$ SOV was significant (by subjects, $\alpha < .01$, and by items, $\alpha < .05$), as well as the preference SVO $>$ OSV ($\alpha < .01$). The low acceptability of the two final orders was expected, as they violate three constraints (VERBALIGN, DOUBLEALIGN, and GROUNDALIGN). There were no other significant differences.

4.6.5.3. Constraint Types

GROUNDALIGN violations caused only mild unacceptability, which is characteristic of soft constraints. Furthermore, GROUNDALIGN seem to be a context-dependent constraint (as defined in Section 4.1.1). The effect of GROUNDALIGN is stronger in the V focus context than in the S focus and O focus context, see Figures 4.17–4.19. (Recall that the constraints is vacuously

satisfied in broad contexts, i.e., in the all focus and null context.) This context effect confirms the status of GROUNDALIGN as a soft constraint.

Like GROUNDALIGN, VERBALIGN seems to induce only mild unacceptability, and might qualify as a soft constraint. However, no clear context effects were found for VERBALIGN; the same relative unacceptability for verb final orders was observed in all contexts (see Figures 4.15–4.19).

DOUBLEALIGN can be classified as a soft constraint, based on the fact that its overall effect on acceptability was weak. Also, DOUBLEALIGN was found to be context-dependent; it caused relatively strong acceptability effects in the all focus context and the V focus context, but led to only small acceptability differences in the null context, S focus context, and O focus context.

4.6.6. Discussion

4.6.6.1. Constraints out of Context

The experimental data provide evidence for the constraints VERBALIGN and DOUBLEALIGN, which are part of our account of the interaction of syntax, phonology, and Information Structure (see (4.47) for the full constraint set). VERBALIGN predicts that verb final orders are reduced in acceptability. This was confirmed by the fact that SOV and OSV found to be consistently dispreferred. DOUBLEALIGN penalizes non-clitic doubled preverbal objects, and is violated by OVS, SOV, and OSV. These orders were less acceptable than SVO, VSO, and VOS, which satisfy DOUBLEALIGN.

4.6.6.2. Constraints in Context

The acceptability patterns found in the context condition were in line with the predictions of the constraint VERBALIGN: the acceptability of verb final orders was reduced. Furthermore, the context condition replicated the results regarding DOUBLEALIGN that were obtained in the null context.

The predictions of the constraint GROUNDALIGN were also born out. GROUNDALIGN requires ground constituents to be sentence peripheral. The effect of this constraint is evident in the S focus context, where VSO was more acceptable than VOS, which violates GROUNDALIGN. In the O focus and V focus context, VSO violates GROUNDALIGN and was less acceptable than VOS (see Figures 4.16–4.19 for details).

We found an unexpected effect involving the verb initial orders VSO and VOS. The experimental data show that the acceptability of these orders is generally reduced compared to SVO. This holds even when the verb initial orders incur no constraint violations and thus are predicted to be as acceptable as SVO. This is an unexpected result in view of the set of constraints in (4.47), and it is unclear how this finding can be explained. However, it seems

unlikely that an explanation in terms of Information Structure is possible. As will be shown below, the effect disappears in Experiment 12. This might be due to the fact that Experiment 12 used speech stimuli, while Experiment 11 was based on written stimuli. As the written language is typically associated with a more formal register, it seems plausible to assume that written stimuli trigger a more normative behavior in the subjects. This would explain the preference for SVO over verb initial orders, as SVO is typically assumed to be the "correct" word order in prescriptive grammars of Greek.

Another result of Experiment 11 concerns the null context condition: here, we found the same pattern as in the all focus context (see Figures 4.15 and 4.16). This is an important methodological finding, as it indicates that even when faced with isolated sentences (which have traditionally been the focus of syntactic research), native speakers make implicit assumptions about Information Structure—they assume an all focus context. We will include the same null context condition in Experiment 12 to test the generality of this result.

4.6.6.3. Constraint Types

We predicted that GROUNDALIGN is a soft constraint, based on the previous results on German obtained in Experiment 10 and on the hypothesis that crosslinguistic variation cannot affect the type of a constraint (see Section 4.1.2). This prediction was confirmed in the present experiment, where the effect of GROUNDALIGN was found to be weak and context-dependent.

The constraint type of VERBALIGN is less clear. On the one hand, a VERBALIGN violation triggers only mild unacceptability, which is typical of soft constraints. On the other hand, no clear context effects could be established for VERBALIGN. We will return to this issue in the modeling study based on the present experiment in Chapter 6, Section 7.6.

For the constraint DOUBLEALIGN clear context effects were found, and we classified DOUBLEALIGN as a soft constraint. Note, however, that no clitic doubled stimuli were included in the present experiment, preventing a full assessment of the effects of DOUBLEALIGN: we cannot check if the clitic doubled version of OVS is really as acceptable as SVO (which is what DOUBLEALIGN predicts). Perhaps OVS is inherently less acceptable than SVO, even under clitic doubling. We will return to this point in Experiment 12, which includes clitic doubled stimuli.

4.6.7. Conclusions

The present experiment provided additional evidence for two important aspects of the distinction between soft and hard constraints proposed in this thesis. Firstly, the experimental findings are consistent with the hypothesis that soft constraint violations are context-dependent, while hard constraints are context-independent. We showed that this hypothesis explains the behavior of GROUNDALIGN and DOUBLEALIGN, which both triggered mild unacceptability and were

subject to context effects, and hence can be regarded as soft constraints.

The experimental results were also compatible with our second hypothesis about the soft/hard dichotomy (see Section 4.1.2): crosslinguistic variation cannot affect the type of a constraint, i.e., no constraints can be soft in one language, but hard in another language. This is in line with the finding that GROUNDALIGN is soft in both German and Greek.

4.7. Experiment 12: Effect of Clitic Doubling, Accent, and Context on Word Order

Experiment 12 is designed to provide further support for the hypothesis that context effects can serve as a diagnostic for constraint types; soft constraints are context-dependent, while hard constraints are context-independent. Furthermore, the results of this experiment will contribute to the understanding of crosslinguistic variation in word order preferences, building on the results on German and Greek in Experiments 6, 10, and 11. These crosslinguistic data will provide the basis for a set of modeling studies in Chapter 6. Note that the present experiment will use spoken instead of written stimuli; we will therefore be able to investigate phonological constraints on word order.

4.7.1. Introduction

Experiment 12 is designed to answer two main questions. Firstly, it investigates the basic claim that clitic doubling and accent placement play an information structural role in a free word order language like Greek. Secondly, the experiment extends the results of Experiment 11 by investigating the validity of a total of five constraints: the word order constraint GROUNDALIGN, the clitic doubling constraints DOUBLEALIGN and DOUBLEGROUND, and the accent constraints ACCENTALIGN and ACCENTFOCUS (see (4.47) for details).

Experiment 12 employs a full factorial design involving the following factors: word order (*Ord*), clitic doubling (*Dou*), accent placement (*Acc*), and context (*Con*). In order to keep the design at a manageable size, only three word orders were included: SVO, OVS, VSO. The order VOS behaved essentially symmetric to VSO in Experiment 11, and was therefore excluded from the present design. The verb final orders were also excluded, as they were mainly used to establish the validity of VERBALIGN, and hence are not essential for the present experiment.

The factor *Dou* had two levels: clitic doubled object and non-doubled object. The following examples represent the clitic doubled versions of the example stimuli in (4.48):

(4.50) a. **SclVO:** O Tasos tha tin diavasi tin efimerida.
the Tasos-NOM will it-CL read-3SG the newspaper-ACC
"Tasos will read the newspaper."
b. **OclVS:** Tin efimerida tha tin diavasi o Tasos.
c. **clVSO:** Tha tin diavasi o Tasos tin efimerida.

The accent factor *Acc* also had two levels: accent on the subject, and accent on the object; consider the following examples:

(4.51) a. **Svo:** O TASOS tha diavasi tin efimerida.
b. **svO:** O Tasos tha diavasi tin EFIMERIDA.

We used the same four contexts for factor *Con* as in Experiment 11, illustrated in (4.49). Again a null context was included as a control condition, enabling us to test the hypothesis that isolated sentences are judged like sentences in an all focus context.

To limit the complexity of the experimental design, we did not include a V accent condition. This means that there is no appropriate intonational realization for the V focus context, where accent is preferred on V. However, we still expect the preference profile for V focus to be informative, as it allows us to investigate the behavior of suboptimal accent realizations (S accent and O accent). Furthermore, the V focus condition is necessary for a full comparison of the results of Experiment 12 with the context effects found in Experiment 11.

4.7.2. Predictions

4.7.2.1. Constraints out of Context

A general prediction is that the acceptability of certain orders (such as OVS) will be affected by clitic doubling. Hence an interaction of *Ord* and *Dou* (clitic doubling) should be present. An interaction of *Ord* and *Acc* (accent placement) is also expected: sentence final accent is preferred by ACCENTALIGN, hence some orders will prefer subject accent, while others will prefer object accent. Finally, we predict an interaction of *Dou* and *Acc*. This follows from the unacceptability of accented clitic doubled objects (see Section 4.6.1.2 for details).

Furthermore, the constraints in (4.47) allow us to make detailed predictions about the acceptability of individual orders. If a given structure violates one of the constraints in (4.47), then we predict its acceptability to be reduced compared a structure that does not incur this constraint violation. These predictions can be tested by further investigating the main effect of *Ord* and the pairwise interactions of *Ord*, *Dou*, and *Acc*. Table 4.3 details which effects will be used to test which constraints. Note that the VERBALIGN, requiring verbs not to be right peripheral, is not relevant, as no verb final orders were included in the present experiment. DOUBLEALIGN, which states that preverbal objects have to be clitic doubled, is violated by OVS. OVS is therefore predicted to be dispreferred compared to SVO and VSO. However,

Table 4.3: Main effects and interactions used to test the constraint set, null context condition (Experiment 12)

interaction	constraints
Ord	GROUNDALIGN
Ord/*Dou*	DOUBLEALIGN
Dou/*Acc*	ACCENTFOCUS, DOUBLEGROUND
Ord/*Acc*	ACCENTALIGN

the difference between OVS and SVO/VSO should disappear in clitic doubled orders, where OclVS satisfies DOUBLEALIGN.

Experiment 11 provided evidence for the hypothesis that a null context behaves like an all focus context. Under this assumption, we can derive predictions from the information structural constraints GROUNDALIGN, DOUBLEGROUND, and ACCENTFOCUS by treating the null context as an all focus context. The constraint GROUNDALIGN, which states that ground constituents have to be sentence peripheral, is vacuously satisfied—there are no ground constituents in an all focus context. The same holds for ACCENTFOCUS, which requires accented constituents to be interpreted as focus. All constituents are in focus, i.e., this constraint is always satisfied, no matter what the accent pattern is.

An interesting case is DOUBLEGROUND, which states that clitic doubled objects have to be interpreted as ground. In stimuli with clitic doubling, DOUBLEGROUND imposes an interpretation where the object is ground. However, as discussed in Section 4.6.1.2, an all focus context may accept a wider range of felicitous answers, including answers with doubled objects (see examples (4.41) and (4.42)). Hence we do not expect an effect of DOUBLEGROUND here. We do, however, predict reduced acceptability for stimuli with accented doubled objects: DOUBLEGROUND states that doubled objects are interpreted as ground; ACCENTFOCUS, however, requires accented constituents to be interpreted as focus. This leads to an inherent, context-independent constraint conflict in orders with object accent and clitic doubling, which are, therefore, predicted to be dispreferred over clitic doubled orders with subject accent and all non-clitic doubled orders.

Finally, ACCENTALIGN requires that accented constituents have to be right peripheral. Hence orders with clause final accent are expected to be preferred: thus, svO should be preferred over Svo, ovS over Ovs, and vsO over vSo. Similarly, for stimuli involving clitics, ACCENTALIGN predicts that Sclvo, Oclvs, and clvSo will be reduced in acceptability.

4.7.2.2. Constraints in Context

On a general level, we expect to find the effects involving *Ord*, *Dou*, and *Acc* that were predicted for the null context condition, i.e., we expect the interactions *Ord*/*Dou*, *Dou*/*Acc*, and *Ord*/*Acc*. The second general prediction is that the accent placement and clitic doubling will interact with

Table 4.4: Interactions used to test the constraint set, context condition (Experiment 12)

interaction	constraints
Ord/*Dou*	DOUBLEALIGN
Dou/*Acc*	ACCENTFOCUS, DOUBLEGROUND
Acc/*Con*	ACCENTFOCUS
Dou/*Con*	DOUBLEGROUND
Ord/*Con*	GROUNDALIGN

Information Structure. Hence, we expect interactions of *Acc* and *Con* and of *Dou* and *Con*. In addition, the interaction of *Ord* and *Con* that was detected in Experiment 11 should be present.

As in the null context condition, we can derive more detailed predictions for individual constraint violations based on the set of constraints in (4.47). These predictions can be tested by further investigating the interactions listed above. Table 4.4 details which interactions will be used to test which constraints.

Firstly, we expect to find the effects that were already discussed for the null context condition: the constraint DOUBLEALIGN is violated in preverbal objects without doubling, i.e., we should find OclVS $>$ OVS.[12] Note, though, that there is the inherent conflict between DOUBLEGROUND and ACCENTFOCUS in stimuli with accented doubled objects which are therefore predicted to be less acceptable than doubled orders with subject accent, and than orders without doubling. As in the null context, we also predict an effect of ACCENTALIGN, i.e., orders with clause final accent are expected to be preferred.

The constraints GROUNDALIGN, DOUBLEGROUND, and ACCENTFOCUS formalize the interaction of order, doubling, and accent with Information Structure. The constraint GROUNDALIGN predicts that orders with non-peripheral ground constituents will be reduced in acceptability (see Experiment 11), while DOUBLEGROUND indicates that stimuli with doubled objects that are not part of ground should be dispreferred. ACCENTFOCUS predicts reduced acceptability for accented constituents that are not in focus.

The following predictions about the effects of GROUNDALIGN, DOUBLEGROUND, and ACCENTFOCUS can be made for each context.

All Focus Context The predictions for the all focus context were already discussed in Section 4.7.2.1, based on the assumption that the null context and the all focus context behave in the same way. To recapitulate: no effects of GROUNDALIGN, DOUBLEGROUND, and ACCENTFOCUS are expected, as these constraints are vacuously satisfied in an all focus context.

S Focus Context In the S focus context, the subject is in focus, while the object is part of ground. This means that GROUNDALIGN is satisfied by SVO, OVS, and VSO, and hence all

[12] Recall that we use "$>$" to denote "is more acceptable than".

three orders would be equally acceptable.

DOUBLEGROUND requires that doubled objects have to be interpreted as ground. This constraint is satisfied, as the S focus context marks the object as ground. Hence our constraint set predicts that doubled and non-doubled orders will be equally acceptable.

ACCENTFOCUS requires that accented constituents are interpreted as focus. This requirement is satisfied by orders with S accent, but violated by orders with O accent, because the S focus context specifies the object as ground. Hence we predict that orders with S accent are more acceptable than orders with O accent.

O Focus Context In the O focus context, the object is in focus, while the subject is part of ground. GROUNDALIGN is satisfied by SVO and OVS, but violated by VSO, where the ground constituent (the subject) is not peripheral. Hence we expect VSO to be reduced in acceptability compared to SVO and OVS.

DOUBLEGROUND requires that doubled objects have to be interpreted as ground. This constraint is violated by clitic doubled orders in the O focus context, as the object is focussed. Hence we predict clitic doubled orders to be less acceptable than doubled ones, which do not violate DOUBLEGROUND.

ACCENTFOCUS is met by orders with O accent, but violated by orders with S accent, as the O focus context specifies the subject as ground. Hence orders with O accent are expected to be more acceptable than S accented orders.

V Focus Context In the V focus context, the verb is in focus, while the subject and the object are ground constituents. As discussed in Experiment 11, VSO incurs a violation of GROUNDALIGN, as the subject fails to be peripheral (i.e., appear either clause finally or clause initially). Hence we predict reduced acceptability for VSO compared to SVO and OVS.

No relevant prediction can be derived from ACCENTFOCUS and DOUBLEGROUND. In the V focus context, all orders violate ACCENTFOCUS, as the accent is either on the subject or on the object (recall that V accent was not included in the stimulus set). DOUBLEGROUND, on the other hand, is satisfied by all orders, as the context marks the object as ground.

4.7.2.3. Constraint Types

In Experiment 11 we used constraint strength and contextual variation to establish that the constraints GROUNDALIGN and DOUBLEALIGN are both soft. We expect this finding to be replicated in the present study; both constraints are expected to trigger only mild unacceptability and be context-dependent.

As mentioned in Section 4.6.1.2, word order is highly ambiguous in information structural terms. On the other hand, accent and doubling are unambiguously associated with focus and ground, respectively. We therefore expect that violations of constraints on accent placement and doubling induce stronger effects than violations of word order preferences. This

means that DOUBLEGROUND and ACCENTFOCUS are expected to be hard constraints, while GROUNDALIGN is a soft constraint, as established already in Experiment 11.

There are no clear predictions as to the type of ACCENTALIGN, a phonological constraint that governs default accent placement in Greek. Whether ACCENTALIGN is soft or hard will be determined based on its constraint strength an contextual behavior.

As in Experiment 11, the present experiment does not allow us to determine constraint rankings directly by analyzing single constraint violations, as the set of experimental factors is not a straightforward implementation of the constraint set under investigation (see Section 4.6.3.3). We will therefore postpone the computation of constraint ranks until Chapter 7, where an automatic procedure for ranking argumentation will be applied to compute a constraint hierarchy based on the data from the present experiment.

4.7.3. Method

4.7.3.1. Subjects

Thirty-six native speakers of Greek participated in the experiment. The subjects were international students at the University of Edinburgh, Napier University, and Heriot-Watt University. The experiment was administered in the laboratory and subjects were paid for their participation. It was made sure that subjects were naive, i.e., they were neither linguists or students of linguistics. None of the subjects had previously participated in Experiment 11.

The data of three subjects were excluded because they were bilingual (by self-assessment). The data of one further subject were excluded as she was a speaker of Cypriot Greek. The data of two subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately. The data of one subject was lost due to a technical problem.

This left twenty-nine subjects for analysis. Of these, 11 subjects were male, 18 female; six subjects were left-handed, 23 right-handed. The age of the subjects ranged from 20 to 37 years, the mean was 26.0 years.

All subjects were resident in Edinburgh at the time of the experiment. The overall time they had lived in an English-speaking environment ranged from 6 to 96 months, the mean was 29.1 months.

4.7.3.2. Materials

Training and Practice Materials These were designed in the same way as in Experiment 1.

Test Materials For the experimental items, a full factorial design was used with word order (*Ord*), context (*Con*), clitic doubling (*Dou*), and accent placement (*Acc*) as the factors (see (4.50) and (4.51) for example stimuli). This yielded a total of $Ord \times Con \times Dou \times Acc =$

$3 \times 5 \times 2 \times 2 = 60$ cells. Eight lexicalizations per cell were used, which resulted in a total of 480 stimuli.

A set of 48 fillers was used, designed to cover the whole acceptability range. Twelve items of each of the following four groups were used: no violation, case violation, phrase structure violation, and agreement violation. The set of fillers was balanced so that each word order and accent pattern used in the experimental items occurred equally often in the fillers. The context items for the fillers where also balanced to reflect the proportions in the experimental set. As in the practice phase, a modulus item in the middle of the range was provided (see Appendix B for a list of all experimental materials).

No frequency matching was conducted for the materials in this experiment, as no adequate corpus was available for Greek.

4.7.3.3. Recordings and Pretests

Recordings Practice and test materials were read by a male native speaker of Greek, who was unaware of the purpose of the experiment. The reader received brief training by the experimenters to make sure that he was able to produce the required accent patterns consistently. The experimental items were tape recorded and later sampled using the sound hardware of a Sparc Ultra 10 workstation. The sampling software used was Sun's Audiotool, with the sampling rate set at 8000 Hz. Questions and answers were recorded separately to exclude possible variations in the accent pattern caused by the context preceding a stimulus during recording.

Intelligibility Pretest As the stimuli crucially relied on phonetically deficient elements (clitics), a pretest was carried out to insure that the stimuli were fully intelligible. Two native speakers of Greek were asked to judge the intelligibility of the stimuli. Under experimental conditions, they listened to the stimuli in random order. Each stimulus was presented once and the subject had to repeat it. The experimenter then compared the repetition to a written version of the stimulus. All stimuli that were not repeated correctly by at least one of the subjects were re-recorded and re-tested. The intelligibility pretest included all experimental items (i.e., the full practice and test sets, including contexts and fillers).

Accent Uniformity Pretest As the stimuli crucially relied on accent placement, a pretest was carried out to ensure that the accent patters were uniformly realized in each experimental condition. Two phonetically trained speakers of Greek (one native and one near-native) were asked to judge whether the accent realized in each experimental condition was uniform across items. Under experimental conditions, the subjects listened to each item in each condition as often as they liked. The were told which accent was supposed to be realized in which condition (S or O accent) and had to judge whether one or more items in the condition had diverging accent patterns. These items were then re-recorded and re-tested. The accent uniformity pretest included only the test items (i.e., contexts, fillers, and practice items were not tested).

4.7.3.4. Procedure

Again, magnitude estimation was used as the experimental paradigm. Each subject took part in an experimental session that lasted approximately 45 minutes and consisted of a training phase, a practice phase, and an experimental phase. The experiment was self-paced, though response times were recorded to allow the data to be screened for anomalies.

The experiment was conducted in a laboratory on PCs. Netscape 4.0 under Windows 95 was used to administer the experiment. The browser established an Internet connection to the experimental server, which controlled the experiment using WebExp 2.1 (Keller et al. 1998).

Instructions We used a Greek version of the instructions in Experiment 1, adapted for spoken stimuli. Where contextualized stimuli were presented, subjects were told that each sentence would be presented in context, defined as a single sentence preceding the target sentence. Subjects were instructed to judge the acceptability of the target sentence, and to take the context into account in their judgments. The task was illustrated by examples.

Demographic Questionnaire and Training Phase These were designed in the same way as in Experiment 1.

Practice Phase This phase allowed subjects to practice magnitude estimation of linguistic acceptability using spoken stimuli. Items were presented to subjects over headphones. For each item, the subject had to click on a Play button to start the presentation of this item. After the item finished playing, the subject had to provide a numeric judgment over the computer keyboard. After pressing Return, the a new Play button for the next item was displayed. Each item had to be played exactly once, and there was no possibility to change responses once Return had been pressed. No time limit was set for the responses.

Subjects first judged the modulus item, and then all the items in the training set. Items were presented in random order, with a new randomization being generated for each subject.

Experimental Phase Presentation and response procedures in the experimental phase were the same as in Experiment 1.

A between-subjects design was used to administer the experimental stimuli: subjects in Group A judged non-contextualized stimuli, while subjects in Group B judged contextualized stimuli.

For Group A, two test sets were used: each set contained four lexicalizations for each of the cells in the design $Ord \times Dou \times Acc$, i.e., a total of 48 items. For Group B, eight test sets were used: each set contained one lexicalization for each of the cells in the design $Ord \times Con \times Dou \times Acc$, a total of 48 items. Lexicalizations were assigned to test sets using Latin squares. Two separate Latin squares were applied: one for the null context condition and one for the context condition.

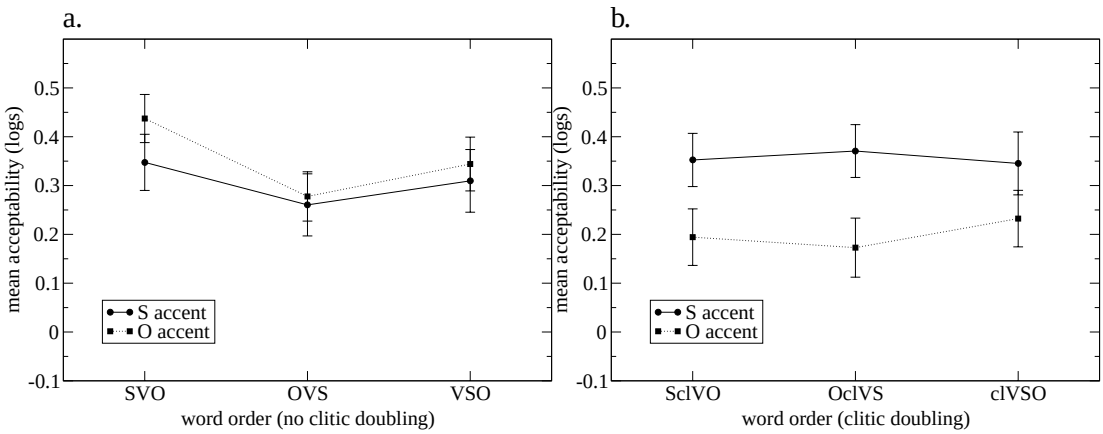

Figure 4.20: Mean judgments for each word order in the null context (Experiment 12)

Subjects first judged the modulus item, which was the same for all subjects. Then they listened to 96 test items: 48 experimental items and 48 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each subject was randomly assigned to a group and a test set; 12 subjects were assigned to Group A, 17 to Group B. Instructions, examples, training items, and fillers were adapted for Group B to take context into account.

4.7.4. Results

The data were normalized as in Experiment 1 and separate ANOVAs were conducted for each subexperiment.

4.7.4.1. Constraints out of Context

The mean judgments for the null context condition are graphed in Figure 4.20. An ANOVA revealed a significant main effect of word order ($F_1(2,22) = 11.873$, $p < .0005$; $F_2(2,14) = 13.704$, $p = .001$). Significant main effects of clitic doubling ($F_1(1,11) = 13.874$, $p = .003$; $F_2(1,7) = 24.555$, $p = .002$) and accent placement were also present ($F_1(1,11) = 10.809$, $p = .007$; $F_2(1,7) = 19.196$, $p = .003$).

As predicted, an interaction between word order and clitic doubling was found ($F_1(2,22) = 7.005$, $p = .004$; $F_2(2,14) = 15.771$, $p < .0005$), indicating that clitic doubling affects the acceptability of certain word orders. We also found an interaction between clitic doubling and accent ($F_1(1,11) = 27.697$, $p < .0005$; $F_2(1,7) = 46.720$, $p < .0005$). This interaction was predicted on the basis of the unacceptability of accented clitic doubled objects. Finally, there was an interaction of word order and accent ($F_1(2,22) = 5.333$, $p = .013$; $F_2(2,14) = 4.442$, $p = .032$). This is in line with the prediction that some word orders prefer S accent, while others prefer O accent. The three-way interaction of word order, clitic doubling,

and accent placement failed to be significant.

Post-hoc Tukey tests were carried out on the interactions to test the predictions of individual constraints, in line with the schema in Table 4.3. The Tukey test for the *Ord*/*Dou* interaction allows us to assess the validity of DOUBLEALIGN, which predicts that OVS (violating DOUBLEALIGN) should be less acceptable than SVO and VSO, while all clitic doubled orders should be equally acceptable. This prediction was borne out: OVS was significantly less acceptable than SVO ($\alpha < .01$) and VSO (by items only, $\alpha < .01$). The Tukey test also showed that SVO was more acceptable than VSO (by items only, $\alpha < .01$), which was unexpected. On the other hand, the orders OclVS, SclVO, clVSO, were not significantly different from each other, in line with our predictions.

It is worth noting that the results in Figure 4.20 support our formulation of DOUBLEALIGN. The theoretical literature on Greek associates the requirement that preverbal objects should be doubled only with ground (unaccented) objects Tsimpli (1995); Tsiplakou (1998). No such restriction is assumed for focused preverbal objects. In contrast, our formulation of DOUBLEALIGN does not make any reference to the discourse function of the preverbal object. If this constraint were to apply only on ground preverbal objects, then Ovs should be much better than ovS, contrary to the results shown in Figure 4.20, where ovS and Ovs receive the same rating.

Furthermore, we predicted reduced acceptability for orders with object accent and clitic doubling, as these orders incur a conflict of DOUBLEGROUND and ACCENTFOCUS. This can be tested by performing a Tukey test on the *Dou*/*Acc* interaction. As predicted, we found that orders with O accent and doubling were significantly less acceptable than orders with S accent and doubling ($\alpha < .01$), orders with S accent without doubling ($\alpha < .01$), and orders with O accent without doubling ($\alpha < .01$). As expected, there were no significant differences between non-doubled orders with S accent, non-doubled orders with O accent, and doubled orders with S accent.

Finally, we conducted a Tukey test on the *Ord*/*Acc* interaction to validate the constraint ACCENTALIGN, which requires that accented constituents have to be right peripheral. This predicts that svO should be preferred over Svo, ovS over Ovs, and vsO over vSo. The Tukey test showed that the preference ovS $>$ Ovs was significant ($\alpha < .01$), but failed to find a difference between svO and Svo, and between vsO and vSo.

4.7.4.2. Constraints in Context

The mean judgments for the context condition are graphed in Figures 4.21–4.24. The ANOVA for the context condition yielded the same general picture as in the non-context condition: significant main effects of word order ($F_1(2,32) = 11.420$, $p < .0005$; $F_2(2,14) = 8.273$, $p = .004$) and clitic doubling ($F_1(1,16) = 20.716$, $p < .0005$; $F_2(1,7) = 17.012$, $p = .004$) were found. Accent, however, failed to reach significance. A main effect of context was also discovered

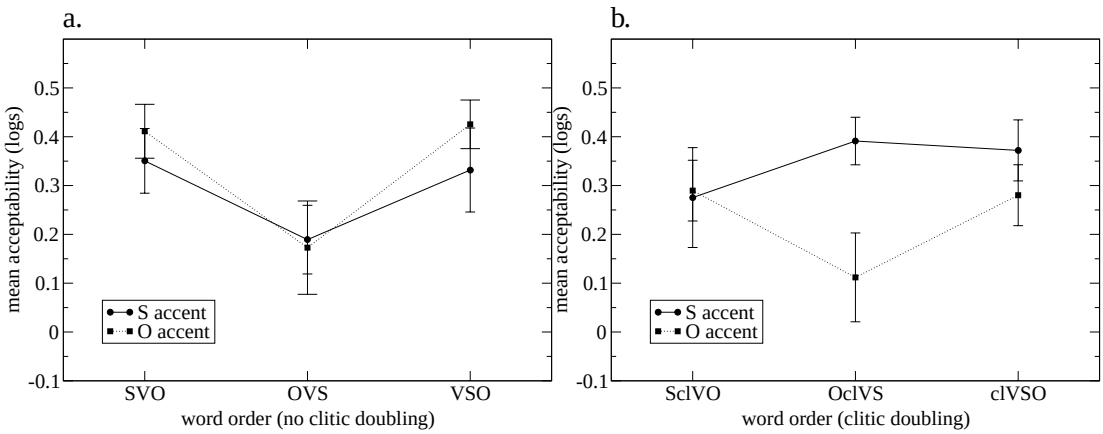

Figure 4.21: Mean judgments for each word order in the all focus context (Experiment 12)

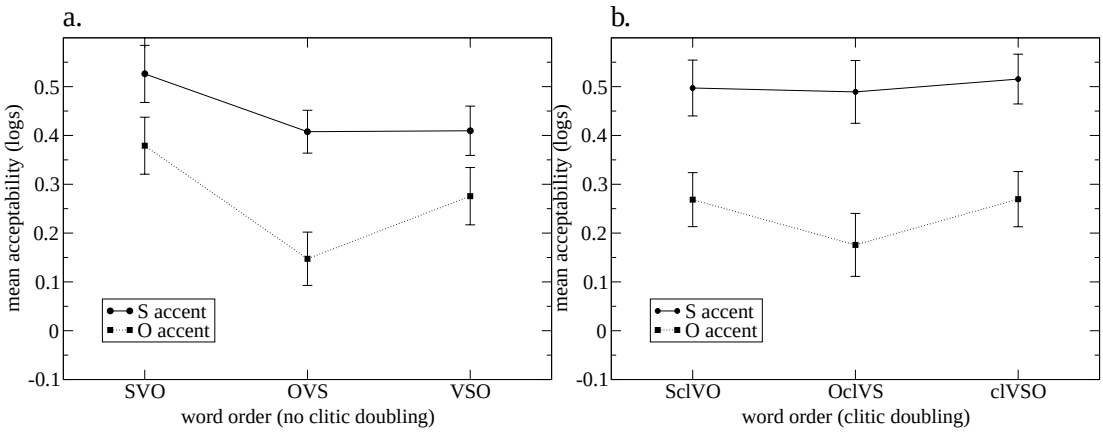

Figure 4.22: Mean judgments for each word order in the S focus context (Experiment 12)

($F_1(3,48) = 11.552$, $p < .0005$; $F_2(3,21) = 28.779$, $p < .0005$).

As in the null context condition, we found an interaction of word order and clitic doubling ($F_1(2,32) = 6.882$, $p = .003$; $F_2(2,14) = 11.565$, $p = .001$), clitic doubling and accent ($F_1(1,16) = 23.439$, $p < .0005$; $F_2(1,7) = 24.133$, $p = .002$), and word order and accent ($F_1(2,32) = 6.284$, $p = .005$; $F_2(2,14) = 5.202$, $p = .020$).

The ANOVA also demonstrated an interaction of accent and context ($F_1(3,48) = 26.359$, $p < .0005$; $F_2(3,21) = 33.098$, $p < .0005$), showing that accent placement has an information structural effect, as predicted. We also discovered an interaction of clitic doubling and context ($F_1(3,48) = 15.155$, $p < .0005$; $F_2(3,21) = 10.869$, $p < .0005$), which confirms that clitic doubling interacts with Information Structure. In addition, we found a significant interaction of word order and context ($F_1(6,96) = 7.722$, $p < .0005$; $F_2(6,42) = 7.124$, $p < .0005$). This confirms the finding in Experiment 11 that word order preferences are subject to context effects. All other interactions failed to reach significance.

Post-hoc Tukey tests were carried out on the interactions to test the predictions of

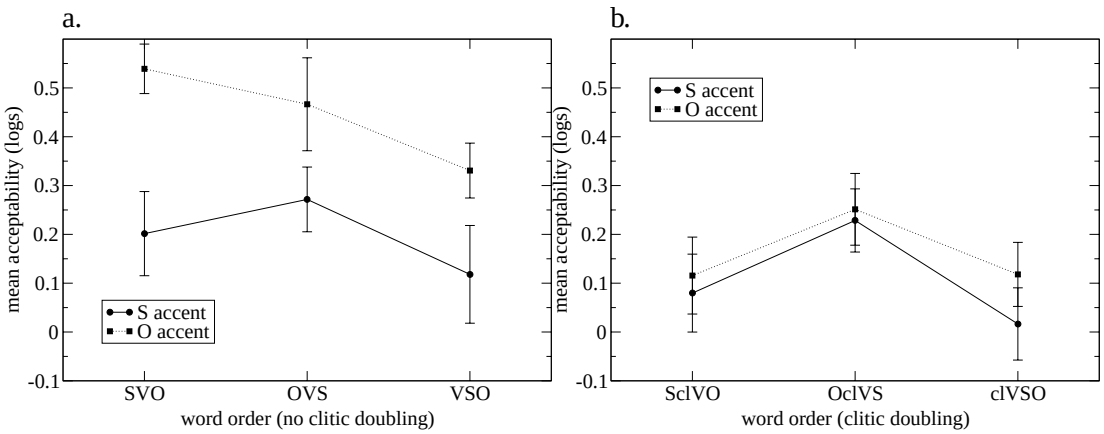

Figure 4.23: Mean judgments for each word order in the O focus context (Experiment 12)

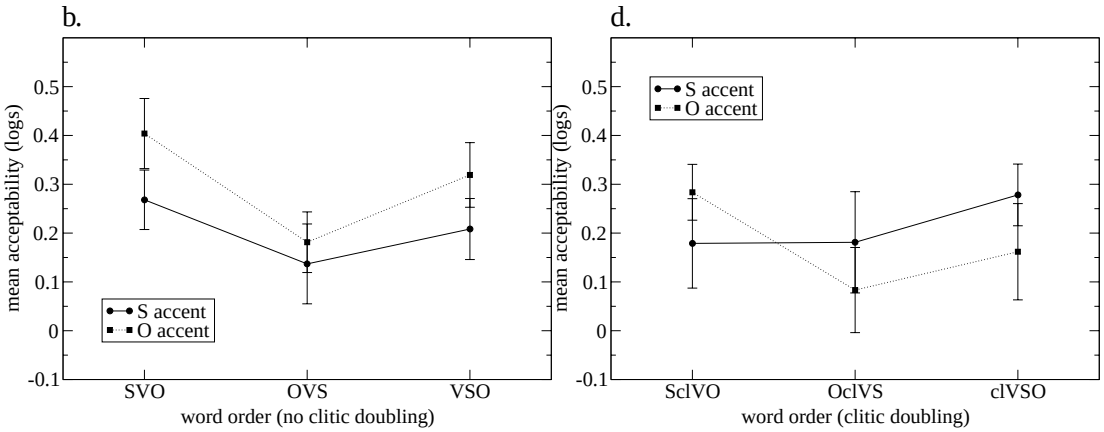

Figure 4.24: Mean judgments for each word order in the V focus context (Experiment 12)

individual constraints (see Table 4.4 for details). We will first report the results for the non-information structural constraints. A Tukey test on the *Ord*/*Dou* interaction was conducted to test the constraint DOUBLEALIGN. As predicted, OVS was less acceptable than SVO ($\alpha < .01$) and VSO (by items only, $\alpha < .05$). The three clitic doubled orders SclVO, OclVS, and clVSO did not differ significantly in acceptability, which is also in line with predictions.

The second context independent prediction was that orders with object accent and clitic doubling should be less acceptable than other orders, as they incur a conflict of DOUBLEGROUND and ACCENTFOCUS. As in the null context condition, we performed a Tukey test on the *Dou*/*Acc* interaction to test this prediction. We found that orders with O accent and doubling were significantly less acceptable than orders with S accent and doubling ($\alpha < .01$), orders with S accent without doubling ($\alpha < .01$), and orders with O accent without doubling ($\alpha < .01$). As expected, there were no significant differences between non-doubled orders with S accent, non-doubled orders with O accent, and doubled orders with S accent.

The constraints GROUNDALIGN, DOUBLEGROUND, ACCENTFOCUS, and

ACCENTALIGN make specific predictions for each context, which we discuss separately below.

All Focus Context In this context, GROUNDALIGN is vacuously satisfied. Therefore we predicted that there should be no difference between the orders SVO, OVS, and VSO. To verify this prediction, we conducted a post-hoc test on the interaction *Ord*/*Con*. There was no significant difference between SVO and VSO, but we found that OVS was significantly less acceptable than both SVO (by items only, $\alpha < .05$) and VSO ($\alpha < .01$), contrary to what was expected. Figure 4.21a provides an explanation for this finding: OVS without doubling violates the constraint DOUBLEALIGN, which greatly reduces its acceptability. This effect is not present in clitic doubled stimuli (see Figure 4.21b).

As mentioned earlier (Section 4.6.1.2), an all focus context can accommodate a wider range of Information Structures. In particular, doubled objects, characteristically associated with a ground interpretation, are felicitous in an all focus context (see examples 4.41 and 4.42 and the relevant discussion in Section 4.6.1.2). Hence DOUBLEGROUND was expected to induce no effects in the all focus context. The Tukey test on the interaction *Dou*/*Con* confirmed this by failing to indicate a significant difference between doubled and non-doubled orders.

ACCENTALIGN predicted that orders with the accent on the rightmost constituent are preferred. We used planned comparisons to test this prediction (post-hoc tests could not be performed as there was no three-way interaction *Acc*/*Con*/*Ord*). Adjusting the significance level using the Bonferroni method, we set $p = .017$, as three comparisons were carried out.

According to ACCENTALIGN, svO should be preferred over Svo, ovS over Ovs, and vsO over vSo. A set of one-way ANOVAs showed that the preference ovS $>$ Ovs was significant (by items only, $F_1(1,17) = 4.74,\ p = .045$; $F_2(1,7) = 20.17,\ p = .003$), but failed to find a difference between svO and Svo, and between vsO and vSo. These results mirrors the ones obtained in the null context, and constitute a partial confirmation of ACCENTALIGN.

S Focus Context In the S focus context, all three orders, SVO, OVS, and VSO, satisfy the constraint GROUNDALIGN and are, therefore, expected to show no significant differences in acceptability. This prediction was born out by a Tukey test on the interaction *Ord*/*Con* (see Figure 4.22).

DOUBLEGROUND requires that doubled objects have to be ground. This requirement is satisfied in an S focus context, where objects are marked as ground elements. Hence doubled and non-doubled orders should be equally acceptable. In line with this prediction, the Tukey test on the interaction *Dou*/*Con* failed to find a significant difference between doubled and non-doubled orders.

The constraint ACCENTFOCUS requires that accented constituents have to be in focus. For the S focus context, this predicts that orders with S accent should be more acceptable than orders with O accent. A Tukey test on the *Acc*/*Con* interaction confirmed this expectation

($\alpha < .01$) (see also Figure 4.22).

Note that there seems to be no effect of ACCENTALIGN in the S focus context.

O Focus Context In this context, GROUNDALIGN is satisfied by SVO and OVS, but violated by VSO. Hence VSO should be reduced in acceptability compared to the verb medial orders (see Figure 4.23). A Tukey test on the *Ord*/*Con* interaction confirmed that VSO was less acceptable than OVS ($\alpha < .01$). The SVO $>$ VSO preference, however, failed to reach significance.

In O focus, orders with clitic doubling violate DOUBLEGROUND and hence are predicted to be less acceptable than non-doubled orders. The Tukey test on the interaction *Dou*/*Con* confirmed this prediction ($\alpha < .01$).

In the O focus context, the constraint ACCENTFOCUS predicts that orders with O accent should be more acceptable than orders with S accent. This prediction was borne out by the Tukey test on the *Acc*/*Con* interaction ($\alpha < .01$) (see also Figure 4.23).

There seems to be no effect of ACCENTALIGN in the O focus context (just like in the S focus context).

V Focus Context In this context, GROUNDALIGN predicts reduced acceptability for VSO compared to SVO and OVS. This prediction could not be confirmed by the Tukey test on the *Ord*/*Con* interaction, which failed to find a difference between VSO and SVO, and between VSO and OVS. However, we found the significant preference SVO $>$ OVS ($\alpha < .01$). This is probably due to the fact that OVS (without doubling) violates DOUBLEALIGN (see also Figure 4.24).

DOUBLEGROUND predicts that doubled and non-doubled orders are equally acceptable as the object is part of ground in the V focus context. In line with this, the Tukey test on the *Dou*/*Con* interaction failed to find a difference between doubled and non-doubled orders.

Note that in the V focus context, ACCENTFOCUS is always violated (as V accent was not included in our stimulus set). This explains why all orders receive fairly low acceptability scores compared to the optimal orders in the O focus and S focus contexts. Furthermore, it seems that the overall acceptability pattern is fairly similar to the one obtained in the all focus context (compare Figures 4.21 and 4.24).

As in the other narrow focus contexts, there was no evidence for an effect of ACCENTALIGN in the V focus context.

4.7.4.3. Constraint Types

As in Experiment 11, the constraint GROUNDALIGN showed a behavior characteristic of a soft constraint: a GROUNDALIGN violation led to mild unacceptability and its effects were context dependent. We found a clear effect of GROUNDALIGN in the O focus context, while

the GROUNDALIGN effect in the V focus context was rather weak (the constraint was not applicable to any other contexts).

DOUBLEALIGN was also found to be context-dependent; it caused strong acceptability effects in the all focus context and the V focus context, but led to only small acceptability differences in the null context, the S focus context, and the O focus context. This is in line with the findings of Experiment 11, where we already concluded that DOUBLEALIGN is a soft constraint. The small overall effect caused by DOUBLEALIGN is also in line with DOUBLEALIGN's status as a soft constraint.

On the other hand, we found that the constraints DOUBLEGROUND and ACCENTFOCUS caused a high degree of unacceptability when violated, which is characteristic of hard constraints. Furthermore, the effect of DOUBLEGROUND and ACCENTFOCUS did not vary with context, which consistent with their status has hard constraints. It was also observed that the interaction of DOUBLEGROUND and ACCENTFOCUS leads to serious unacceptability in all stimuli with accented doubled objects (where the two constraints are inherently in conflict). This effect was very general: it applied to all contexts, including the null context and the all focus context (where DOUBLEGROUND or ACCENTFOCUS does not apply on its own). Based on the absence of context effects, we therefore conclude that both DOUBLEGROUND and ACCENTFOCUS are hard constraints.

The constraint on ACCENTALIGN, which requires accent to fall on the rightmost constituent was found to be context-dependent; it only triggers acceptability effects in broad contexts (null context and all focus context). Note that the effect of a ACCENTALIGN is weak, which also suggest that we are dealing with a soft constraint.

4.7.5. Discussion

4.7.5.1. Constraints out of Context

Experiment 12 provided evidence for the constraint DOUBLEALIGN. DOUBLEALIGN requires preverbal objects to be doubled and is satisfied by SVO and VSO, but violated by OVS. The experimental findings in the null context condition were in line with these predictions: for non-doubled orders, SVO and VSO were significantly more acceptable than OVS. The doubled orders OclVS, SclVO, and clVSO, on the other hand, did not differ in acceptability, as predicted by DOUBLEALIGN. These results extend the findings of Experiment 11, that only tested DOUBLEALIGN on non-doubled stimuli.

The null context condition also provided evidence for an interaction of the two information structural constraints DOUBLEGROUND and ACCENTFOCUS. DOUBLEGROUND requires doubled objects to be ground, while ACCENTFOCUS requires accented constituents to be focused. The two requirements are in conflict for accented doubled objects, which would have to be ground and focus at the same time. This is an inherent conflict that does not depend

on the focus-ground structure imposed by the context, hence accented doubled objects should be unacceptable in all contexts. This prediction that was born out in the null context.

We also tested ACCENTALIGN, which requires the accent to fall on the rightmost constituent. This prediction was partly borne out; ovS was found to be more acceptable than Ovs; but we failed to find a difference between svO and Svo, and between vsO and vSo.

4.7.5.2. Constraints in Context

The context condition confirmed the results for DOUBLEALIGN obtained in the null context condition. It also replicated the interaction of DOUBLEGROUND and ACCENTFOCUS obtained in the null context: accented doubled objects were unacceptable in all contexts, in line with the prediction that the conflict between DOUBLEGROUND and ACCENTFOCUS is context-independent.

Furthermore, the context condition allowed us to test the constraint GROUNDALIGN which requires ground constituents to be sentence peripheral. This prediction was borne out in the S focus context, where SVO, OVS, and VSO were not significantly different. In the O focus context, we found that VSO, which contains a non-peripheral ground subject, was less acceptable than OVS, as predicted by GROUNDALIGN. These results are in line with the findings regarding GROUNDALIGN obtained in Experiment 11.

We also tested the predictions of DOUBLEGROUND, the constraints that states that doubled objects have to be interpreted as ground. This constraint is satisfied in the all focus, S focus, and V focus context. Doubled and non-doubled stimuli were equally acceptable in these contexts, as predicted. In O focus, doubled stimuli violate DOUBLEGROUND and were less acceptable than non-doubled ones. The constraint ACCENTFOCUS requires that accented constituents have to be interpreted as focus; this was confirmed in the S focus context, where stimuli with S accent were more acceptable than stimuli with O accent. In the O focus context, the pattern was reversed. ACCENTALIGN predicts that accented constituents have to be right peripheral. Tendencies in line with the predictions of ACCENTALIGN could be observed in the all focus context.

To summarize, the present experiment extended the results of Experiment 11 by providing evidence for a total of five grammatical constraints: the word order constraints GROUNDALIGN, the clitic doubling constraints DOUBLEALIGN and DOUBLEGROUND, and the accent constraints ACCENTALIGN and ACCENTFOCUS (see (4.47) for details). All of these constraints were well supported by the experimental findings, with the exception of ACCENTALIGN, which only manifested itself in weak tendencies. Further experimental data will be necessary to back up ACCENTALIGN.

Another important finding of Experiment 12 is that an all focus context behaves like a null context (compare Figure 4.20 and Figure 4.21). This replicates the results of Experiment 11 for a wider range of context sensitive phenomena and for spoken stimuli, thus providing further

support for the hypothesis that subjects make minimal contextual assumptions when they are exposed to isolated sentences: a null context is treated like an all focus context, which is what is expected under an information structural approach.

4.7.5.3. Constraint Types

As mentioned in Section 4.6.1.2, word order is highly ambiguous in information structural terms. On the other hand, accent and doubling are unambiguously associated with focus and ground, respectively. We therefore predicted that violations of constraints on accent placement and doubling induce stronger effects than violations of word order preferences. This prediction is in line with the experimental results, which suggested that DOUBLEGROUND and ACCENTFOCUS are hard constraints, while GROUNDALIGN is a soft constraint. Furthermore, the experimental findings confirm the status of DOUBLEALIGN as a soft constraint (in line with Experiment 11), and also establish that ACCENTALIGN is a soft constraint—its effects were weak and context dependent.

Taken together, the results from Experiment 11 and 12 allow us to draw a clear distinction between soft constraints like GROUNDALIGN, DOUBLEALIGN, and ACCENTALIGN that are context-dependent and trigger only mild unacceptability, and hard constraints like ACCENTFOCUS and DOUBLEGROUND that are context-independent and trigger serious unacceptability. This finding provides clear support for claim that constraint strength and context effects can serve as diagnostic for the type of a constraint.

Note that the results of the present experiment fail to provide additional evidence for the claim (see Section 4.1.2) that crosslinguistic variation cannot alter the type of a constraint. Only the constraint GROUNDALIGN was tested crosslinguistically; its status as a soft constraint in both German and Greek was already established in Experiments 10 and 11. However, the phonological constraint ACCENTFOCUS seems to be a good candidate for a constraint that is crosslinguistically hard; it is plausible to assume that a violation of this constraint leads to serious unacceptability in all languages that use accent to mark focus. Also the constraints VERBINITIAL and VERBFINAL (not tested in the present experiment) can be assumed to be crosslinguistically hard. We will provide evidence for this in a modeling study involving subordinate clauses in Greek in Chapter 6, Section 7.4.

4.7.6. Conclusions

The results of this experiment provided additional support for the main hypothesis of the present chapter. They demonstrated for a set of five constraints that soft constraints are context-dependent, while hard constraints are context-independent: a violation of a hard constraint causes the same degree of unacceptability in all contexts, while the opposite was true for soft constraints. By investigating spoken stimuli, the present experiment significantly expanded the

range of data supporting this claim: the inclusion of spoken stimuli enabled us to test the context hypothesis for a set of phonological constraints.

The present experiment makes interesting predictions with respect to the second hypothesis that was central to the present chapter: that the type of a constraint (hard or soft) does not vary from language to language. DOUBLEGROUND and ACCENTFOCUS were found to be hard constraints. This leads us to expect that these constraints are hard constraints also in other language, a prediction can be tested by determining if DOUBLEGROUND and ACCENTFOCUS lead to strong unacceptability and context-independent effects in languages other than Greek. Along the same lines, we predict that the constraints DOUBLEALIGN and ACCENTALIGN should be soft across languages. Investigating these predictions will be left to further research.

4.8. Conclusions

The present experiment expanded the investigation of extraction and word order that we began in Chapter 3. It also presented experimental data on a further phenomenon, gapping. This completes the experimental part of the present thesis (except for the methodological studies in Chapter 5). The set of linguistic phenomena considered in Chapters 3 and 4 was designed to cover all the major grammar modules standardly assumed in syntactic theory (see Section 3.1.5 for an overview). Such a design allows us to make maximally general claims about the behavior of gradient linguistic structures.

The data reported in this chapter re-iterated a main point of the preceding chapter: gradient acceptability judgments (collected experimentally) allow us to settle data disputes in theoretical linguistics. This was evidenced by our findings on gapping (Experiments 7 and 8), extraction (Experiment 9), and word order (Experiments 10–12). For example, we were able to show that gapping is equally acceptable with adjunct and complement remnants, a fact that is controversial in the theoretical literature. Also, we demonstrated that the acceptability of pronominalized orders in German is context-dependent, which is not predicted by existing accounts. Furthermore, we provided evidence regarding the acceptability of preverbal clitic-doubled objects in Greek, traditionally the subject of data disputes in the theoretical literature. This confirms that such data disputes are the results of the informal data collection techniques employed in theoretical linguistics, which are not well-suited to investigate the behavior of gradient linguistic data.

In Chapter 3, we derived a number of general properties of gradient linguistic data. These properties concern the classification of constraints into types, and the ranking and interaction of constraints. Regarding constraint types, the experimental data reported in the present experiment are compatible with the hypothesis that constraints cluster into two types, soft and hard constraints, based on the following set of criteria:

- **Gradience** In Chapter 3, we found that soft constraint violations are associated with

mild unacceptability, while hard violations trigger serious unacceptability. This finding was broadly consistent with the experimental data reported in the present chapter. However, certain soft and hard violations can trigger a similar degree of unacceptability (see Experiment 8). This indicates that a constraint cannot be classified as hard or soft based solely on its constraint rank: additional criteria (context effects and crosslinguistic effects) have to be taken into account.

- **Context Effects** The main focus of this chapter was the hypothesis that soft constraints are context-dependent, while hard constraints are context-independent. The experimental results we presented provided a wealth of evidence of this hypothesis, leading to the claim that context effects can serve as a diagnostic for the type of a constraint.

- **Crosslinguistic Variation** Based on data on crosslinguistic variation in word order preferences, we were able to investigate the claim (advanced in Chapter 3) that crosslinguistic effects are limited to soft constraints. We replaced this claim by the more accurate hypothesis that both hard and soft constraints are subject to crosslinguistic variation (constraint re-ranking), but that crosslinguistic variation cannot affect the type of a constraint (i.e., there are no constraints that are soft in one language and hard in another).

The main results regarding constraint ranking and constraint interaction were already established in Chapter 3. However, the experimental data reported in the present chapter expanded the empirical base of the relevant claims:

- **Ranking** Both soft and hard constraints are ranked, i.e., constraints can differ in the degree of unacceptability triggered by a constraint violation. We confirmed this claim by establishing constraint hierarchies for gapping (in Experiment 8) and word order (in Experiment 10).

- **Cumulativity** Constraint violations are cumulative, i.e., the unacceptability of a structure increases with the number of constraints it violates. Chapter 3 had already provided robust evidence for this hypothesis, which could be confirmed in Experiments 8 and 10.

- **Strict Domination** Experiment 10 provided further evidence for the ganging up of constraint violations, and against OT-style strict domination of constraint. IT showed that soft constraint can gang up against hard ones, consistent with Experiment 5.

Taken together, the experimental results in Chapters 3 and 4 provide a wealth of information about the properties of gradient linguistic judgments. Chapter 6 will develop a model of gradient grammaticality that accounts for these properties. Chapter 6 will also discuss other models

of gradience proposed in the literature, and evaluate them against the data presented in Chapters 3 and 4. In Chapter 7, we will then test our model of gradience by providing detailed accounts of the extraction data obtained in Chapter 3, as well as modeling the gapping and word order data presented in Chapter 4.

Before we proceed to these two theoretical chapters, we will turn to a number of methodological considerations that relate to the web-based experimental paradigm used throughout this thesis. The next chapter will discuss the reliability and validity of this paradigm.

Chapter 5

Methodological Aspects

Most of the experimental results presented in Chapters 3 and 4 were obtained using a web-based experimental methodology. While this mode of experimentation allows rapid access to a large number of subjects (even for less commonly spoken languages), it raises important questions as to its reliability and validity compared to more conventional experimental methodologies.

The present chapter addresses these questions. We first discuss the problems and opportunities that arise from web-based experimentation and explain the safeguards that were put in place for the experiments reported in this thesis. We then present a number of experiments that demonstrate the reliability and validity of web-based studies. This includes the web-based replication of the results of a lab-based study and a questionnaire-based study.

5.1. Introduction

Most of the experiments discussed in Chapters 3 and 4 were administered using the World-Wide Web, a method that has proved controversial in the recent experimental literature (e.g., Johnson-Laird and Savary 1999; Mehler 1999). It has been argued that by using web data, the experimenter can exercise less control over the experimental setting, as each subject might complete the experiment under different conditions, possibly in an environment that includes noise or other distractions. Also, there is an obvious need for making sure that the subjects taking part in the experiment respond in the way intended by the experimenter, i.e., that they understand and follow the experimental instructions properly. A third problem is subject authentication—we have to guarantee that the subject provides genuine data and does not take part more than once in each experiment.

In this section, we discuss how these problems are addressed by the software used to administer Experiments 1–12.

5.1.1. Experimental Procedure

Experiments 1–12 were administered using WebExp (Keller et al. 1998), a software package designed for conducting psycholinguistic studies over the web (for general recommendations on Internet experiments see Hewson, Laurent, and Vogel 1996).[1]

WebExp is implemented as a set of Java classes. As Java is a full-fledged programming language, it gives the web designer maximal control over the interactive features of a web site. WebExp makes use of this flexibility to keep the experimental procedure as constant as possible across subjects. An important aspect is that the sequence in which the experimental items are administered is fixed for each subject: the subject does not have the ability to go back to previous stimuli or to inspect or change previous responses. (If the subject hits the "back" button on the browser, the experiment will terminate.)

Another important feature is that WebExp provides precise timings of subject responses by measuring the response onset time and the completion time for each answer (with an accuracy of approximately 60ms). These timings are useful in screening the responses for anomalies, i.e., to eliminate the subjects who responded too quickly (and thus probably did not complete the experiment in a serious fashion), or those who responded too slowly (and thus probably were distracted while doing the experiment). WebExp automatically tests the response timings against upper and lower limits provided by the experimenter and excludes subjects whose timings are anomalous. Further manual checks can be carried out on the response timings later on.

5.1.2. Subject Authentication

Apart from providing response timing, WebExp also offers a set of safeguards that are meant to ensure the authenticity of the subjects taking part, and exclude subjects from participating more than once.

1. **Email address** Each subject has to provide their email address. An automatic plausibility check is conducted on the address to ensure that it is syntactically valid. If the address is valid, then WebExp automatically sends an email to this address (containing a message thanking the subject for taking part). If the email bounces, the experimenter should exclude this subject from the data set, as they probably used a fake identity.

2. **Personal data** Before being allowed to start the experiment, each subject has to fill in a short questionnaire supplying name, age, sex, handedness, and language background. These data allow manual plausibility checks to be conducted, and subjects that give implausible answers can be eliminated from the data set.

[1] For more information on WebExp, see `http://www.hcrc.ed.ac.uk/web_exp/`. Experiments using WebExp can be accessed through a central entry point at `http://surf.to/experiments/`.

3. **Responses** A manual inspection of the responses allows us to detect subjects that have misunderstood the instructions and responded in an anomalous fashion, e.g., by giving the same response to every item.

4. **Connection data** The software also logs the following data related to the subject's web connection: Internet address of their machine, operating system and browser they use, and the URL from which they accessed the experiment (the referring web page). This information (in addition to the email address) is valuable in detecting subjects that take part more than once.

Note that taking part in a WebExp study requires a subject to give up their anonymity and supply name and email address. This is a move we consider justified in the interest of ensuring subject authenticity. The experimental web site contains a privacy statement that guarantees that all subject data will be treated strictly confidential.

5.2. Experiment 13: Reliability of Web-based Experiments

5.2.1. Introduction

The safeguards outlined in Section 5.1 go some way towards ensuring that our web-based methodology is sound. However, to provide a rigorous evaluation of web-based data, we need to prove the reliability and validity of the experimental procedure used.

The present experiment tests the *reliability* of the web-based procedure by comparing the results of two web-based studies carried out on the same materials.[2] The *validity* of the web-based procedure can be established by comparing web data to data obtained using conventional questionnaire-based or lab-based methods. Such comparisons are reported in Sections 5.3 and 5.4, respectively.

The present replication study deals with gradient acceptability in extraction from picture NPs. A subset of the materials of Experiment 4 was used for the replication: we included only those materials containing soft constraint violations, i.e., violations of the constraints on definiteness, referentiality, and verb class. Also, the set of fillers differed between the original study and the replication. The experimental protocol and the subject population from which we sampled were identical in both experiments.

5.2.2. Predictions

Our hypothesis is that there is no difference between the response patterns obtained in Experiment 4 and its replication. If this hypothesis is correct, then the same significant effects should

[2] Note that we are not strictly speaking establishing test-retest reliability, as two distinct samples of subjects were used. This is common practice in psycholinguistics, where learning effects can be expected if the same subject is tested on the same materials more than once.

be obtained for both data sets. Furthermore, we can perform an ANOVA on the combined data, treating the experimental condition (original or replication) as a between-groups factor. Under the hypothesis that there is no difference between the two data sets, a main effect of experimental condition, and in particular, interactions between experimental condition and the other factors should be absent. Finally, we can test the hypothesis that there is a linear relationship between the judgments obtained by the two studies by performing a correlation analysis on the two data sets.

5.2.3. Method

5.2.3.1. Subjects

Twenty-nine native Speakers of English from the same population as in Experiment 4 participated in the experiment. None of the subjects had previously participated in Experiment 4, 5, or 9.

The data of two subjects were excluded because they were linguists (by self-assessment). The data of two subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 25 subjects for analysis. Of these, 17 subjects were male, eight female; four subjects were left-handed, 21 right-handed. The age of the subjects ranged from 17 to 72 years, the mean was 35.0 years.

5.2.3.2. Materials

Training and practice materials were designed in the same way as in Experiment 1.

The test materials used where the subset of the materials of Experiment 4 that dealt with soft constraint violations (involving the factors *Def*, *Ref*, and *Verb*). This subset was chosen as it was felt that replicating the effects from soft violations is more difficult (and thus provides a stronger form of validation) than replicating the effects from hard violations, where the size of the effect is much larger (see Section 3.5.5).

The experimental design was $\mathit{Def} \times \mathit{Ref} \times \mathit{Verb} = 2 \times 2 \times 2$, yielding a total of eight cells. For each cell, the four lexicalizations were used that were also employed in Experiment 4, yielding 32 stimuli in total. A set of 16 fillers was used, designed to cover the whole acceptability range.

5.2.3.3. Procedure

Four test sets were used: each test set contained one lexicalization for each of the 16 cells in the design. Lexicalizations were assigned to test sets using a Latin square covering the full set of items.

Subjects first judged the modulus item, which was the same for all subjects and remained on the screen all the time. Then they saw 32 test items: 16 experimental items and 16 fillers. Items were presented in random order, with a new randomization being generated for each subject. Each experimental subject was randomly assigned to one of the test sets.

The remainder of the experimental procedure used was the same as in Experiment 1.

5.2.4. Results

The data were normalized as in Experiment 1.

For the original data set (taken from Experiment 4), an ANOVA revealed significant main effects of *Verb* ($F_1(1,25) = 17.075$, $p < .0005$; $F_2(1,3) = 17.234$, $p = .025$) and *Ref* ($F_1(1,25) = 14.612$, $p = .001$; $F_2(1,3) = 11.765$, $p = .042$). The effect of *Def* was significant by subjects, and marginal by items ($F_1(1,25) = 8.152$, $p = .009$; $F_2(1,3) = 7.199$, $p = .075$). All interactions failed to be significant.

The ANOVA for the replication study showed the same significant effects. There were main effects of *Verb* ($F_1(1,25) = 12.457$, $p = .002$; $F_2(1,3) = 55.51$, $p = .005$) and *Ref* ($F_1(1,25) = 15.126$, $p = .001$; $F_2(1,3) = 21.694$, $p = .019$). Again, the effect of *Def* was significant by subjects, and marginal by items ($F_1(1,25) = 4.754$, $p = .039$; $F_2(1,3) = 5.722$, $p = .097$). All interactions failed to be significant.

To further test the hypothesis that the original and the replication study yielded the same results, we conducted an ANOVA on the combined data set, treating the experimental condition (original or replication) as a between-groups variable.[3] This ANOVA yielded a main effects of *Verb* ($F_1(1,49) = 28.958$, $p < .0005$; $F_2(1,3) = 71.433$, $p = .003$) and *Ref* ($F_1(1,49) = 29.238$, $p < .0005$; $F_2(1,3) = 24.796$, $p = .016$), and *Def* ($F_1(1,49) = 12.874$, $p = .001$; $F_2(1,3) = 55.701$, $p = .005$). There was no main effect of experimental condition, and all interactions between experimental condition and the other factors were non-significant.

Finally, we conducted a correlation analysis that compared the average judgments for each cell in the two data sets. A highly significant correlation was obtained by subjects and by items ($r_1 = .9024$, $p = .002$, $N = 8$; $r_2 = .9204$, $p = .001$, $N = 8$).

5.2.5. Discussion

We presented a replication of Experiment 4, focusing on the effects of soft violations (factors *Verb*, *Ref*, and *Def*). These effects were chosen because their small effect sizes make them harder to replicate than hard constraint violations, which are typically associated with large effects.

Separate ANOVAs on the original data set and on the data from the replication study revealed the same significant effects. We also failed to find an effect of experimental condition

[3] Note that this ANOVA is not a case of multiple tests on the same data. Rather we refine the two previous ANOVAs by including experimental condition as an additional factor. Hence there is no need to adjust the *p*-value.

(original or replication) in an ANOVA on the pooled data. More importantly, there were no interactions between experimental condition and the other experimental factors. We further showed that there is a high correlation between the average judgments obtained in both experiments.

Taken together, these results amount to a full replication of the results from Experiment 4. This demonstrates that our web-based experimental procedure is reliable, i.e., two samples taken from the same population yield comparable results.

5.3. Experiment 14: Validity of Web-based Experiments against Questionnaire-based Experiments

5.3.1. Introduction

Experiment 13 showed that web-based experiments are reliable, i.e., that carrying out the same experiment on two different samples from the same population yields comparable results. The present experiments carries the validation of web-based data a step further by replicating the results of a questionnaire study from the literature.

For our replication, we chose Gordon and Hendrick's (1997) study on binding theory. Gordon and Hendrick (1997) present a series of experiments that tested native speakers' knowledge of binding principles using a coreference judgment task. In this task, subjects were asked to judge the acceptability of sentences like (5.1), under the assumption that the expressions in boldface refer to the same person. (Note that this methodology is the same that was employed in our Experiment 5.)

(5.1) **She** adores **Zelda's** teachers.

Our replication comprised Experiments 1–4 of Gordon and Hendrick's (1997) study. We briefly outline the design of these four experiments.

Experiment 1 This study was designed to test Principle C of binding theory. Three factors were manipulated: *Ana*, i.e., the type of the NP sequence (name-pronoun, name-name, or pronoun-name); and *Com*, i.e., whether the first noun phrase c-commands the second. The third factor was *Subj*, i.e., whether the antecedent was located in the subject (as in (5.2a)) or in the object (as in (5.2b)). Example stimuli (with the antecedent in the subject) are given in Table 5.1.

(5.2) a. **John's** roommates met **him** at the restaurant.
b. Jane introduced **Bill's** new teacher to **him**.

Experiment 2 In this experiment, Gordon and Hendrick (1997) replicated the results of Experiment 1 for sentences where the antecedent is contained in an adjunct. The factors *Ana* (binding configuration) and *Com* (c-command) where the same for as in Experiment 1, yielding a total of six binding configurations, examples of which can be found in Table 5.2.

Table 5.1: Sample stimuli from Gordon and Hendrick (1997), Experiment 1

NP_1	NP_2	c-command	sample sentence
name	pronoun	no	**John's** roommates met **him** at the restaurant.
name	pronoun	yes	**John** met **his** roommates at the restaurant.
name	name	no	**John's** roommates met **John** at the restaurant.
name	name	yes	**John** met **John's** roommates at the restaurant.
pronoun	name	no	**His** roommates met **John** at the restaurant.
pronoun	name	yes	**He** met **John's** roommates at the restaurant.

Table 5.2: Sample stimuli from Gordon and Hendrick (1997), Experiment 2

NP_1	NP_2	c-command	sample sentence
name	pronoun	no	Before **Susan** began to sing **she** stood up.
name	pronoun	yes	**Susan** stood up before **she** began to sing.
name	name	no	Before **Susan** began to sing **Susan** stood up.
name	name	yes	**Susan** stood up before **Susan** began to sing.
pronoun	name	no	Before **she** began to sing **Susan** stood up.
pronoun	name	yes	**She** stood up before **Susan** began to sing.

Experiment 3 This experiment extended the results of Experiments 1 and 2 by including stimuli containing anaphora (reflexives), thus allowing us to compare the effects of Principles A, B, and C of binding theory. Again, c-command was manipulated in the stimuli, resulting in a total of eight binding configurations. Examples are listed in Table 5.3.

Experiment 4 This experiment elaborated on Experiment 1 by testing the effects of Principle C in two additional configurations: either inside a possessive NP (as in Experiment 1), or inside a conjoined NP (see (5.3)); this was factor *Conj*. Furthermore, the antecedent could either be in the subject (as in (5.2a) and (5.3a)) or in the object (as in (5.2b) and (5.3b)); this was factor *Subj*. The binding configurations tested were the same as in Experiment 1 (see Table 5.1).

(5.3) a. **Jeff** and Cindy asked the bakery to make a cake for **him**

Table 5.3: Sample stimuli from Gordon and Hendrick (1997), Experiment 3

NP_1	NP_2	c-command	sample sentence
name	pronoun	no	(1) **Joan's** father respects **her**.
pronoun	name	no	(2) **Her** father respects **Joan**.
name	name	no	(3) **Joan's** father respects **Joan**.
pronoun	anaphor	no	(4) **Her** father respects **herself**.
name	anaphor	no	(5) **Joan's** father respects **herself**.
name	pronoun	yes	(6) **Joan** respects **her**.
pronoun	name	yes	(7) **She** respects **Joan**.
name	anaphor	yes	(8) **Joan** respects **herself**.

b. Jill told **Dustin** and Sara that **he** was uninsured.

Gordon and Hendrick (1997) used a binary judgment task for their Experiments 1–3 (coreference is possible or not). For Experiment 4 this was modified to an ordinal judgment task: subject had to judge the acceptability of coreference on an ordinal scale from 1 (completely unacceptable) to 6 (completely acceptable). This task is more similar to the magnitude estimation task that we used for our replications studies.

Furthermore, Gordon and Hendrick (1997) used two different sets of instructions in Experiment 4. "Reflective" instructions required subjects to read the stimulus once, repeat it to themselves, and then rate the acceptability of coreference. "Immediate" instructions asked for subjects' initial reaction after having read the stimulus once. Our replication used only one set of instructions, which left open how often subjects should read each stimulus.

5.3.2. Predictions

We do not expect that the web-based study will replicate the results of the questionnaire-based study perfectly. Apart from the difference in administering the experiment (over the web or with a questionnaire in the classroom), there were a number of other differences between Gordon and Hendrick's (1997) original and our replication:

1. Gordon and Hendrick (1997) sampled from a different subject population: they used university students attending an Introduction to Language course. The replication study sampled from the population of English-speaking web users.

2. For Experiments 1–3, Gordon and Hendrick (1997) use a nominal scale (acceptable or unacceptable), while in Experiment 4, they use an ordinal scale with 6 points. For our replication experiments, however, we used magnitude estimation, i.e., an interval scale.

3. Gordon and Hendrick (1997) used relatively large sample sizes (around 45 subjects per experiment), while the replication only used samples of about 15 subjects per experiment.

Due to these differences, we do not predict a perfect match between the original and the replication study. However, we can hypothesize that differences (2) and (3) work in opposite directions: the replication uses a more sensitive measurement scale, and thus should be able to detect acceptability differences with fewer subjects than the original.

In general, we predict that the replication study will find the same significant effects as the original, and that there should be a high correlation between the average judgments in the original data set and in the replication.

5.3.3. Method

5.3.3.1. Subjects

Sixty-eight native Speakers of English from the same population as in Experiment 4 participated in the experiment.

The data of another subject were excluded because he was a linguist (by self-assessment). The data of six subjects were eliminated after an inspection of the responses showed that they had not completed the task adequately.

This left 61 subjects for analysis. Of these, 30 subjects were male, 31 female; six subjects were left-handed, 55 right-handed. The age of the subjects ranged from 17 to 57 years, the mean was 28.4 years.

5.3.3.2. Materials

Training and practice materials were designed in the same way as in Experiment 1.

Experiment 1 Following Gordon and Hendrick (1997), the design was $\mathit{Ana} \times \mathit{Com} \times \mathit{Subj} = 3 \times 2 \times 2$, yielding a total of 12 cells. For each cell two lexicalizations from Gordon and Hendrick (1997) were used, which resulted in a set of 24 items (see Table 5.1 for sample stimuli). The third lexicalization (containing relative clauses) was omitted to keep the size of the stimulus set small.

Experiment 2 Following Gordon and Hendrick (1997), the design was $\mathit{Ana} \times \mathit{Com} = 3 \times 2$, yielding a total of six cells. For each cell the four lexicalizations from Gordon and Hendrick (1997), which resulted in a set of 24 items (see Table 5.2 for sample stimuli).

Experiment 3 Following Gordon and Hendrick (1997), the design contained only the factor *Ana* with 8 levels. Three lexicalizations were used. One was the original lexicalization used by Gordon and Hendrick (1997), the other two were new lexicalizations, similar to the original one. This resulted in a set of 24 items (see Table 5.2 for sample stimuli).

Experiment 4 The design of this experiment was $\mathit{Ana} \times \mathit{Com} \times \mathit{Subj} \times \mathit{Conj} = 3 \times 2 \times 2 \times 2$, yielding a total of 24 cells. However, half of the stimuli (the ones with possessive antecedents) were identical to the ones used in Experiment 1. These stimuli were omitted from the replication, making *Conj* a between-groups factors and reducing the size of the design to 12 cells. For each cell, the two lexicalizations from Gordon and Hendrick (1997) were used, which resulted in a set of 24 items.

5.3.3.3. Procedure

Each subject was randomly assigned to one of the four stimulus sets and judged 24 experimental items and 24 fillers. 16 subjects took part in Experiment 1, and 15 each in Experiments 2–4.

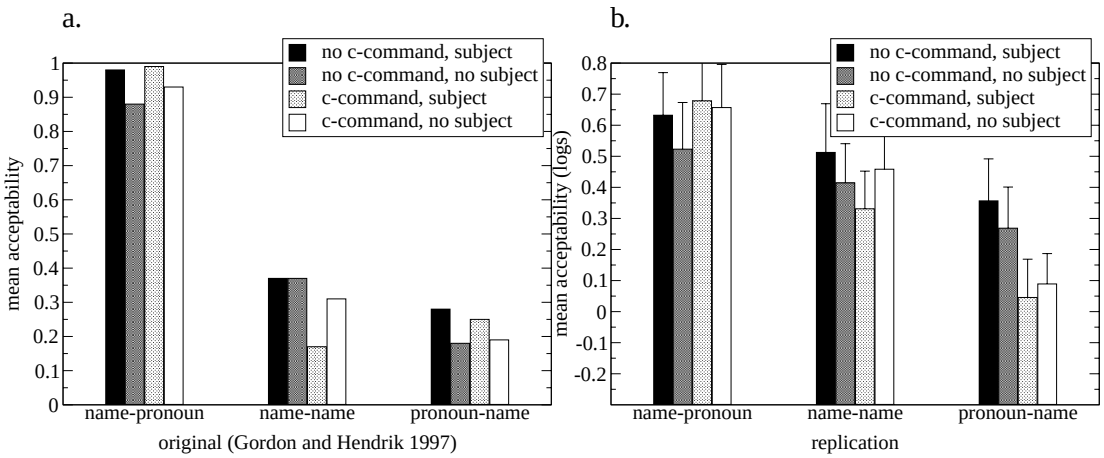

Figure 5.1: Replication of Gordon and Hendrick (1997), Experiment 1

The remainder of the experimental procedure was the same as in Experiment 1.

5.3.4. Results

The data were normalized as in Experiment 1.

Experiment 1 The average judgments for the different conditions are graphed in Figure 5.1 for both the original study and our validation study. Visual inspection of the data shows that the patterns for the four conditions of the name-pronoun configuration and name-name configuration are replicated well in the validation study.

Gordon and Hendrick (1997) found a significant main effect of *Ana*, i.e., of the type of NP sequence. This effect was replicated in our study ($F_1(2,30) = 16.799$, $p < .0005$).[4] Gordon and Hendrick (1997) also found a weak main effect of *Com*, i.e., of c-command. This effect failed to be present in our data. Finally, Gordon and Hendrick (1997) reported an interaction between *Ana* and *Com*, which could be replicated ($F_1(2,30) = 6.189$, $p = .006$).

To determine the locus of the interaction of *Ana* and *Com*, Gordon and Hendrick (1997) conducted post-hoc *t*-tests, adjusted by the Bonferroni method. They found a significant effect of c-command in the name-name configuration, but not for the name-pronoun and pronoun-name configuration.

In our replication study, we also conducted a series of post-hoc tests to further probe the interactions, adjusting for multiple comparisons using the Bonferroni method.[5] As three comparisons were carried out, we set $p = 0.016$ as our significance level. We found a marginal

[4]Gordon and Hendrick (1997) do not report F_2 values, probably because their experiments use a small number of lexicalizations (typically 2 or 3).

[5]Gordon and Hendrick (1997) are not explicit about their Bonferroni adjustments, i.e., they do not specify the number of comparisons they assume, and only report adjusted *p*-values. In the following, we will explicitly provide this information, based on our reconstruction of their experimental design. All *p*-values have to be interpreted relative to the adjusted significance level we specify for each experiment.

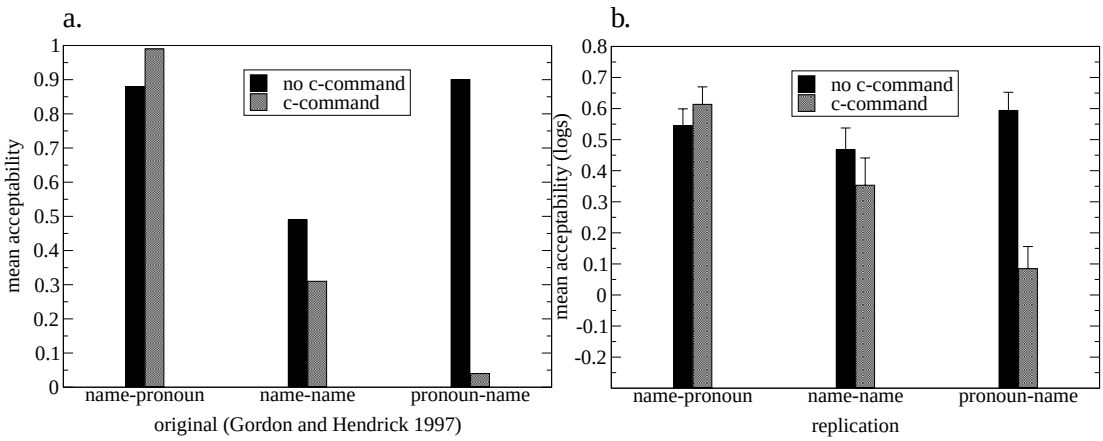

Figure 5.2: Replication of Gordon and Hendrick (1997), Experiment 2

difference for the name-pronoun condition ($t_1(15) = 2.619$, $p = .019$), while the difference in the name-name and pronoun-name conditions failed to be significant.

Furthermore, Gordon and Hendrick (1997) report an interaction between type of antecedent (in the subject or not) with c-command. This interaction was also present in our data ($F_1(1,15) = 5.436$, $p = .034$). In the original study, a post-hoc *t*-test demonstrated that the locus of this effect was the name-name sequence: c-command had a significant effect if the antecedent was in the subject, but was not significant if the antecedent was outside the subject. We conducted to post-hoc *t*-tests an replicated this results: there was a significant effect of c-command for name-name sequences if the antecedents was in the subject ($t_1(15) = 2.711$, $p = .016$), we found no effect if the antecedent was in the object. (This assumes a Bonferroni adjustment for two comparisons, i.e., $p = .025$.)

Experiment 2 The average judgments for the different conditions are graphed in Figure 5.2 for both the original study and our replication. Again, the replication study mirrors the acceptability pattern for each of the binding configurations.

Gordon and Hendrick (1997) found a significant main effect of *Ana*, i.e., of the type of NP sequence. This effect was replicated in our data ($F_1(2,28) = 12.888$, $p < .0005$). Furthermore, Gordon and Hendrick (1997) report a main effect of *Com*, i.e., of c-command relationship, which was also attested in the replication ($F_1(1,14) = 20.886$, $p < .0005$). Finally, an interaction of *Ana* and *Com* was found both in the original and in the replication study ($F_1(2,28) = 28.434$, $p < .0005$).

To determine the locus of the effect of c-command, Gordon and Hendrick (1997) conducted post-hoc *t*-tests. They found a significant effect of c-command for all three binding configurations. We replicated these tests and adjusted for multiple comparisons using the Bonferroni methods, i.e., we set the significance level at $p = .0167$, as three comparisons were carried out. We found a significant effect of c-command for the name-pronoun condition

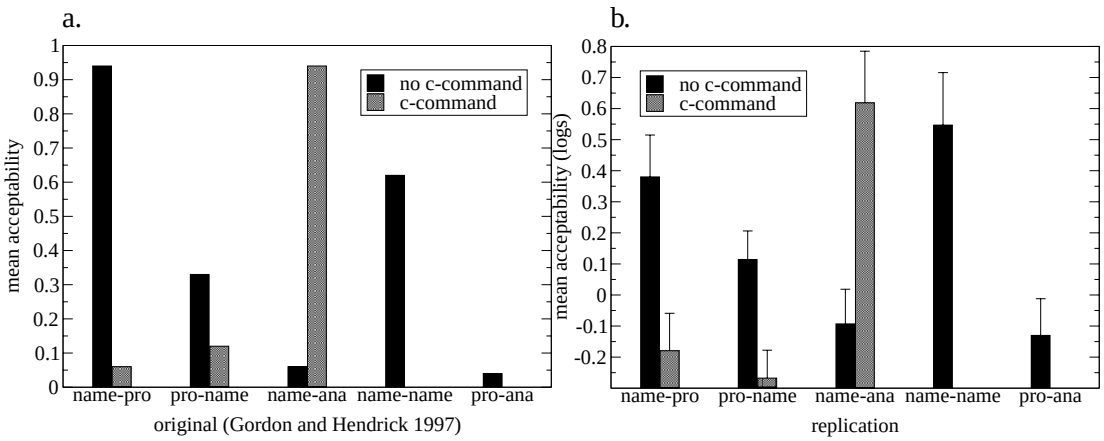

Figure 5.3: Replication of Gordon and Hendrick (1997), Experiment 3

($t_1(14) = 2.979$, $p = .010$) and for the pronoun-name condition ($t_1(14) = 6.411$, $p < .0005$), but not for the name-name condition.

Experiment 3 The average judgments for the different conditions are graphed in Figure 5.3 for both the original study and our replication. Visual inspection shows that the replication experiment produces the same acceptability pattern for each of the binding configurations.

This was confirmed by the statistical analyses. Gordon and Hendrick (1997) report a significant main effect of sentence type, which was also present in our data ($F_1(7,98) = 17.561$, $p < .0005$). They also found that the acceptability of the name-anaphor configuration increases under c-command, which was replicated in our data ($F_1(1,14) = 17.057$, $p = .001$). Another finding was that c-command significantly reduces acceptability in name-pronoun configurations. This effects was also present in the replication ($F_1(1,14) = 21.818$, $p < .0005$). An effect of c-command on the acceptability of pronoun-name configurations was also found both in the original data set and in our replication ($F_1(1,14) = 25.949$, $p < .0005$). Finally, a comparison of the name-pronoun and the name-name configurations showed that names are favored as antecedents ($F_1(1,14) = 13.770$, $p < .002$), in line with what Gordon and Hendrick (1997) found.

Experiment 4 The average judgments for the different conditions are graphed in Figures 5.4 and 5.5 for both the original study and our replication. As in Experiments 1–3, the acceptability patterns obtained for each binding configuration in the replication study are highly similar to the pattern in the original experiment.

We computed an ANOVA on the combined data from Experiment 1 and Experiment 4 to evaluate this experiment (recall that in the replication, the possessive antecedent condition was shared between the two experiments). This means that the factor *Conj* (possessive or conjoined antecedent) was a between-groups factor in our replication, while it was a within-groups in the original study. Another difference between the replication and the original was that the

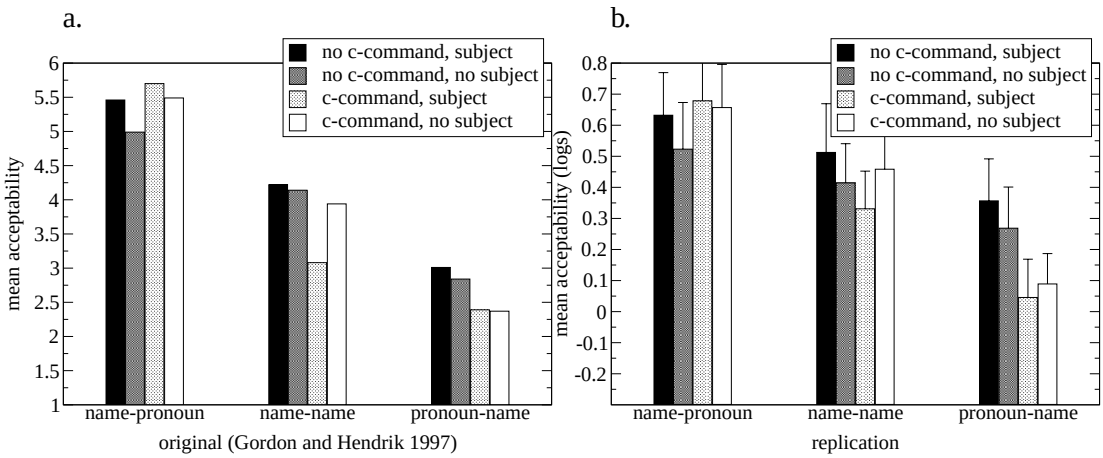

Figure 5.4: Replication of Gordon and Hendrick (1997), Experiment 4, possessive antecedent

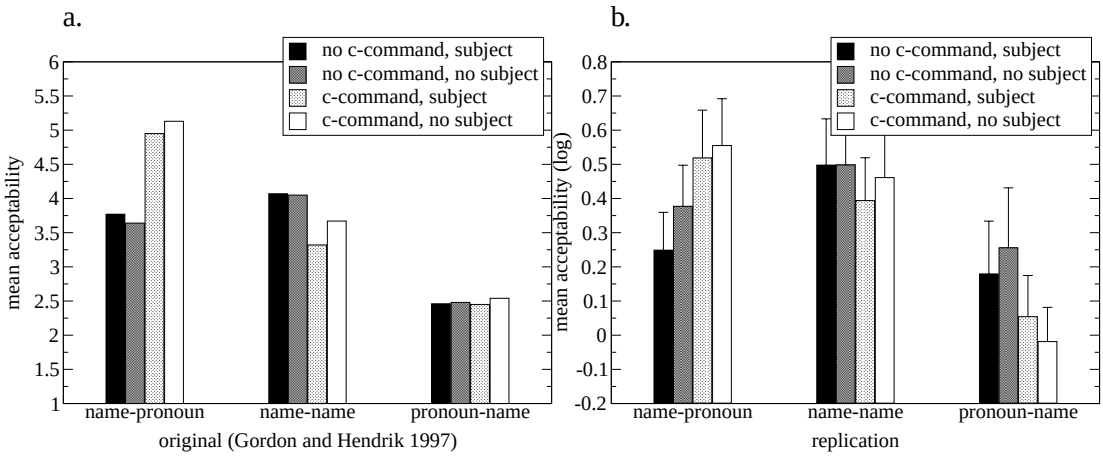

Figure 5.5: Replication of Gordon and Hendrick (1997), Experiment 4, conjoined antecedent

replication study used only one set of instructions, while the original used two. (All the comparisons reported below were carried out on the averages for the two instruction sets.) As a consequence of these differences in experimental design, the results of the original ANOVA and the replication ANOVA are not strictly comparable, even though the acceptability patterns in both studies are very similar (see Figures 5.4 and 5.5).

Gordon and Hendrick (1997) failed to find a main effect of *Com* (c-command). This effect was also absent in the replication study. The significant main effect of *Ana* (type of NP sequence) reported by Gordon and Hendrick (1997) could be replicated ($F_1(2,58) = 30.185$, $p < .0005$). Also, there was a significant interaction between *Com* and *Ana* ($F_1(2,58) = 12.703$, $p < .0005$), just as in the original study. Post-hoc test were conducted for the interaction of *Com* and *Ana*. The significance level was adjusted using the Bonferroni method for three comparisons, i.e., we set $p = .0167$. C-command had a significant effect on acceptability in the name-pronoun condition ($t_1(30) = 4.580$, $p < .0005$), the name-name condition ($t_1(30) = 2.727$,

$p = .011$), and in the pronoun-name condition ($t_1(30) = 2.727$, $p = .009$). All three effects were also present in the original study.

Like Gordon and Hendrick (1997) we failed to find a main effect of *Subj*, i.e., subject or object antecedent. Furthermore, Gordon and Hendrick (1997) report an interaction of *Subj* and *Com*, an interaction of *Subj* and *Ana*, and a three-way interaction of *Subj*, *Com*, and *Ana*. None of these interactions was present in the replication data. However, separate ANOVAs on the replication data for possessive antecedents (shared with Experiment 1) and conjoined antecedents (particular to Experiment 4) showed that the interaction of *Subj* and *Com* was present in the possessive data ($F_1(1,15) = 5.436$, $p = .034$), while the interaction *Subj*, *Com*, and *Ana* was present in the conjoined data ($F_1(2,28) = 4.343$, $p = .023$). The fact that these interaction did not show up in the overall ANOVA was probably due to the reduced power of the between-groups design used in the replication study.

The original study reports the results of post-hoc contrasts on these interactions. We replicated these contrasts, again adjusting the significance level according to the Bonferroni methods, i.e., setting $p = .0125$ for four comparisons. The original experiment demonstrated that in the name-name condition, c-command had a significant effect for both subject antecedents and object antecedents. The replication study only found an effect of c-command for subject antecedents ($t_1(30) = 3.783$, $p = .001$). For the pronoun-name condition, the original study produced a significant effect of c-command for subject antecedents, and a marginal one for object antecedents. The replication revealed marginal effects in both cases ($t_1(30) = 2.350$, $p = .026$ and $t_1(30) = 2.529$, $p = .017$).

The original study found a main effect of type of antecedent NP (possessive or conjoined, factor *Conj*), which was absent in our data. Furthermore, the original study found an interactions between *Conj* and *Ana*, which was marginal in the replication ($F_1(2,28) = 2.648$, $p = .079$). An interaction between *Conj* and *Com*, as well as a three-way interaction *Conj*/*Com*/*Ana* was also present in the original. These interactions could not be replicated.

To further analyze the *Conj*/*Com*/*Ana* interaction, Gordon and Hendrick (1997) conducted a series of post-hoc tests, which we replicated, setting the significance level to $p = .0083$ (Bonferroni adjustment for six comparisons). For the name-pronoun condition, Gordon and Hendrick (1997) found an effect of c-command for both possessive and conjoined antecedents. In our data, we only found an effect of c-command for conjoined antecedents ($t_1(14) = 4.059$, $p = .001$). In the original, effects of c-command for both possessive and conjoined antecedents were also found for the name-name condition and the pronoun-name condition; these four effects failed to reach significance in the replication study.

Correlations To further compare the results of the original experiments and our validation study, we conducted correlation analyses comparing the mean judgments for each cell in the experiment. High correlations were obtained for Experiment 2 ($r_1 = .9050$, $p < .0005$, $N = 24$) and Experiment 3 ($r_1 = .9198$, $p = .001$, $N = 8$). The correlation for Experiment 1 was

somewhat lower ($r_1 = .7567$, $p < .0005$, $N = 24$).

For Experiment 4, we carried out separate analyses for the two instruction conditions used by Gordon and Hendrick (1997) (reflective vs. intuitive instructions). A high correlation was obtained in the reflective condition ($r_1 = .9009$, $p < .0005$, $N = 48$), while the correlation in the intuitive condition was somewhat lower ($r_1 = .8436$, $p < .0005$, $N = 48$). The correlation of the two conditions in the original data was high as well ($r_1 = .9650$, $p < .0005$, $N = 48$).

5.3.5. Discussion

In Experiments 2 and 3, we achieved a full replication of the original experiments, i.e., all significant main effects, interaction, and post-hoc contrasts reported by Gordon and Hendrick (1997) were attested in our data. In Experiments 1 and 4, where the experimental design was more complicated, the main results of the original study were replicated, with some interactions and post-hoc contrast failing to reach significance. We obtained high correlations between the original data and the replication data for all four experiments.

Taken together, these results amount to a successful replication of Gordon and Hendrick (1997) results. This is particularly remarkable given that other factors than the experimental procedure (questionnaire vs. web) differed between the original study and our experiment (see Section 5.3.2): we sampled from a different subject population (web users), an interval scale was used for subjects' responses, and only small samples were utilized.

An interesting difference between the original and our replication concerns the name-name binding configuration. In Experiment 1, Gordon and Hendrick (1997) obtained relatively low judgments for this configuration, similar to the judgments for the pronoun-name configuration (see Figure 5.1a). On the other hand, in Experiment 4, which was a partial replication of Experiment 1, they found that the name-name stimuli patterned in between the name-pronoun (highly acceptable) and the pronoun-name (highly unacceptable) configuration (see Figures 5.4a and 5.5a).

In our replication study, however, the name-name stimuli behaved consistently: they patterned between name-pronoun and pronoun-name in Experiment 1 as well as in Experiment 4 (see Figures 5.1b, 5.4b, and 5.5b). This indicates that the behavior of name-name stimuli in Gordon and Hendrick's (1997) Experiment 1 is an artifact of the nominal response scale they used for this experiment. The nominal scale forces the subjects to make a choice between acceptable and unacceptable coreference, which obscures the fact that name-name sequences are of intermediate acceptability. The intermediate status these constructions is uncovered, however, when an ordinal or interval scale is used (as in the original Experiment 4 and our replication of Experiments 1 and 4).

5.4. Experiment 15: Validity of Web-based Experiments against Lab-based Experiments

5.4.1. Introduction

In the preceding sections we presented a reliability study for web-based data and a validation study comparing web-based data to published questionnaire data. The ultimate test for web-based experimentation, however, is a comparison of web-based data against data obtained with a conventional, lab-based procedure (using the same experimental software).

To provide such a comparison, a set of statistical tests were carried out on the data that overlap between Experiment 11 (web-based) and Experiment 12 (lab-based). The overlapping data constitute acceptability judgments for three different word orders (SVO, OVS, VSO) in five contexts (null, all new, S new, O new, V new), yielding a factorial design of $Ord \times Con = 3 \times 5$. For Experiment 12, all analyses were carried out on the average of the judgments for both accent patterns (S and O accent), in order to make this data comparable to the judgments for written stimuli in Experiment 11. (The underlying assumption is that subjects assign an accent at random if they see written stimuli, clearly an idealization.)

5.4.2. Predictions

Our hypothesis is that there is no difference between the response patterns obtained over the web and in a laboratory setting. If this hypothesis is correct, then the same significant effects should be obtained for both data sets. Furthermore, we can perform an ANOVA on the combined data from the web-based and the lab-based studies, treating the experimental procedure as a between-groups factor. Under the hypothesis that there is no difference between the two data sets, a main effect of the experimental procedure, and in particular, interactions between the procedure and the other experimental factors should be absent. Finally, we can test the hypothesis that there is a linear relationship between the web-based and the lab-based judgments by performing a correlation analysis on the two data sets.

5.4.3. Method

The method for Experiment 11 (web-based) and Experiment 12 (lab-based) is described in Sections 4.6.4 and 4.7.3.

5.4.4. Results

Null Context Condition As in Experiments 11 and 12, the non-context condition and the context condition were analyzed separately. In the non-context condition, the ANOVA for the web-based study yielded an effect of *Ord* (word order), which was significant by subjects only

($F_1(2,32) = 14.552$, $p < .0005$; $F_2(2,14) = 3.181$, $p = .156$). An effect of *Ord* was found also in the lab-based study, this time significant both by items and by subjects ($F_1(2,32) = 22.514$, $p < .0005$; $F_2(2,14) = 37.294$, $p < .0005$). A further ANOVA was carried out on the combined data from the web-based and the lab-based study, treating the experimental procedure (web or lab) as a between-groups variable.[6] Again, we found a highly significant effect of word order ($F_1(2,54) = 30.535$, $p < .0005$; $F_2(2,14) = 7.775$, $p = .005$). There was a main effect of *Exp* (experimental procedure), which was significant by items, but non-significant by subjects ($F_1(2,54) = 2.155$, $p = .154$; $F_2(1,7) = 123.109$, $p < .0005$). That *Exp* was significant by items has an obvious explanation in the fact that Experiment 11 used written materials, while Experiment 12 used speech stimuli. As predicted, there was no interaction between *Exp* and word order, indicating that the type of experiment (web or lab) did not interfere with the overall word order effect.

Context Condition The ANOVAs for the context condition confirmed the results for the null context. On the web-based data, we found significant effects of *Ord* (word order) ($F_1(2,32) = 34.678$, $p < .0005$; $F_2(2,14) = 12.246$, $p = .001$) and *Con* (context) ($F_1(3,48) = 3.371$, $p = .026$; $F_2(3,21) = 5.456$, $p = .006$). An interaction of *Ord* and *Con* was also present ($F_1(6,96) = 2.417$, $p = .002$; $F_2(6,42) = 2.417$, $p = .043$). As predicted, the lab-based study showed exactly the same pattern: there was a significant effect of *Ord* ($F_1(2,32) = 19.933$, $p < .0005$; $F_2(2,14) = 9.491$, $p = .002$) and *Con* ($F_1(3,48) = 3.980$, $p = .013$; $F_2(3,21) = 6.096$, $p = .004$), and an interaction of the two factors ($F_1(6,96) = 7.058$, $p < .0005$; $F_2(6,42) = 4.056$, $p = .003$).

We also carried out an ANOVA on the pooled data from the web-based and the lab-based study, using the experimental procedure *Exp* (web or lab) as a between-groups variable.[7] The effect of *Ord*, *Con*, and the interaction *Ord*/*Con* were highly significant. As in the null context condition, there was a main effect of *Exp*, which was significant only by items, ($F_1(1,32) = .810$, $p = .375$; $F_2(1,7) = 12.325$, $p = .01$). Again, this is probably an effect of stimulus type (written or speech). As predicted, there was no interaction of *Exp* with any of the other factors, indicating that the experimental procedure did not affect the overall acceptability pattern.

Finally, we conducted a correlation analysis that compared the judgments for each cell in the web-based and the lab-based data set. For the context condition, a highly significant correlation was obtained by subjects and by items ($r_1 = .895$, $p < .0005$, $N = 12$; $r_2 = .917$, $p < .0005$, $N = 12$). This can be considered strong evidence that the subjects behaved in a similar fashion under both experimental conditions. (A correlation analysis for the null context condition could not be conducted, as the number of data points was too small.)

[6] Note that this ANOVA is not a case of multiple tests on the same data. Rather we refine the two previous ANOVAs by including experimental condition as an additional factor. Hence there is no need to adjust the p-value.

[7] Again, the p-value was not adjusted for multiple comparisons.

5.4.5. Discussion

We presented a re-analysis of the data obtained in Experiments 11 and 12 to back up the claim that web-based experimental data and laboratory data yield comparable results. Separate ANOVAs on both data sets revealed the same significant main effects and interactions. We also failed to find by-subjects effects of experimental procedure in an ANOVA on the pooled data, both for the null context and for the context condition. By-item effects of the experimental procedure were obtained, which can be explained straightforwardly by the fact that the web experiment used written stimuli, while the lab study employed spoken materials. Crucially, there was no interaction between experimental procedure and the other experimental factors. We further showed that there is a high correlation between the average judgments obtained with both procedures. Taken together, these results suggest that there is no difference between web-based and lab-based data, at least as far as the purpose of the present study is concerned

5.5. Conclusions

This chapter dealt with issues relating to conducting psycholinguistic experiments over the web, which was the experimental procedure employed in most of the studies presented in Chapters 3 and 4.

In Section 5.1, we discussed the pros and cons of web-based experimentation and outlined the features of WebExp, the software package used for conducting Experiments 1–12. WebExp is designed to keep the experimental procedure as constant as possible across subjects. It provides precise timing data for each response, which allows us to eliminate subjects that fail to complete the experiment in a serious manner (leading to very high or very low response times). WebExp also records a range of subject data which can be used to screen out bogus subjects, or subjects that participate more than once. These data include personal details, email address, and data relating to the subject's Internet connection.

In Section 5.2, we presented results that demonstrated the reliability of our web-based experimental procedure. We showed that a replication of Experiment 4 yields the same results as in the original (i.e., the same significant effects). We also demonstrated a high correlation between the data in the original study and the replication.

In Sections 5.3 and 5.4 we established the validity of web-based data by comparing it to data obtained using conventional experimental methodologies. Section 5.3 showed that a near-perfect replication of the results of a published questionnaire-based study can be achieved using web data. The majority of the effects reported in the original experiments were replicated and high correlations were obtained between the original data sets and the replication data. In Section 5.4, this results was extended to lab-based data. We re-analyzed the data that overlap between Experiment 11 and Experiment 12, and demonstrated that the web-based and the lab-based experiment yield the same significant effects. Again, the web data and the lab data were

highly correlated.

To summarize, the results presented in this chapter give strong evidence for the claim that web data is as reliable and valid as data obtained with conventional methodologies. It is an open question if this finding extends beyond the type of data used in this thesis (judgments of linguistic acceptability).

Chapter 6

A Model of Gradient Grammaticality

This chapter deals with modeling gradient linguistic data. We first identify a set of criteria that an adequate model of gradience has to meet. Then we discuss previous proposals in the literature on the basis of these criteria and identify their shortcomings.

The problems with existing models of gradient grammaticality prompt us to propose a new model of gradience that borrows central concepts from Optimality Theory, a competition-based grammatical framework. This model, Linear Optimality Theory, is motivated by the experimental results on constraint ranking and the cumulativity of constraint violations in Chapters 3 and 4. The core assumption of Linear Optimality Theory is that linguistic constraints are annotated with numeric weights, and that the grammaticality of a structure is determined by the weighted sum of the constraint violations it incurs. We show that Standard OT (which uses ranks instead of weights) is a special case of Linear Optimality Theory.

The chapter also deals with the problem of estimating the parameters (the constraint weights) for a Linear Optimality Theory model. This problem can be reduced to the problem of solving a system of linear equations, for which standard algorithms exist, such as Gaussian Elimination or Least Square Estimation. These algorithms have attractive computational properties when applied to Linear Optimality Theory.

6.1. Introduction

This section summarizes the main properties of gradient linguistic structures uncovered by the experiments in Chapters 3 and 4. Based on these properties, we outline a set of criteria that can be used to assess models of gradient grammaticality.

6.1.1. Properties of Gradient Linguistic Structures

Based on experimental data covering a range of syntactic phenomena in several languages, a set of general properties of gradient constraint violations could be identified in Chapters 3 and 4.

The two central findings are that constraints are ranked and that constraint violations are cumulative. Constraint ranking means that some constraint violations are significantly more unacceptable than others. Cumulativity means the multiple constraint violations are significantly more unacceptable than single violations. Ranking and cumulativity effects were exhibited by all the constraints investigated in Chapters 3 and 4. These properties seem to be fundamental to the behavior of gradient linguistic judgments and therefore should form the basis of a model of gradience in grammar. Cumulativity also accounts for the ganging up effect that was observed in Experiments 4, 5, and 10: multiple violations of low ranked constraints can be as unacceptable as a single violation of a higher ranked constraint.

Another central experimental result is that constraints can be classified into two types, soft and hard. While both types of constraint share the properties of ranking and cumulativity, they differ in another set of properties (see Table 6.1). First, soft constraint violations are associated with mild unacceptability, while hard violations trigger serious unacceptability. Second, soft constraints are context-dependent, while hard constraints are immune to context effects. Third, although both hard and soft constraints are subject to crosslinguistic variation (constraint re-ranking), crosslinguistic variation cannot affect the type of a constraint (i.e., there are no constraints that are soft in one language and hard in another).

Experiment 4 raised the possibility that there is a limited form of OT-style strict domination, viz., that soft constraints are strictly dominated by hard ones. However, the results of Experiments 5 and 10 allowed us to rule out this possibility by showing that ganging up effects also hold between constraint types, i.e., that multiple soft constraint violations can gang up against a single hard violation.

Based on these three properties, we can operationalize the notion of constraint type. If a constraint violation induces strong unacceptability and fails to show context effects, then it can be classified as a hard constraint. If a constraint triggers only mild unacceptability and is subject to contextual variation, then the constraint is soft. The classification can be verified by investigating the crosslinguistic behavior of the constraint; the type of a constraint (soft or hard) should remain the same across languages.

A model of gradience in grammar such as the one proposed in the present chapter (see Section 6.3) will have to account for the experimental properties of gradient linguistic structures that were summarized in this section. This set of properties will also be a valuable tool in evaluating existing models of gradience (see Section 6.2).

6.1.2. Criteria for Models of Gradient Grammaticality

This section outlines a set of criteria that we will use in the remainder of the chapter to assess models of gradient grammaticality. These criteria are an extension of the ones initially proposed by Keller (1996a: Ch. 4).

Table 6.1: Properties of hard and soft constraints

	hard constraints	soft constraints
universal effects	ranking effects cumulativity effects ganging up effects	ranking effects cumulativity effects ganging up effects
type-specific effects	strong unacceptability no context effects crosslinguistic variation limited to hard constraints	mild unacceptability context effects crosslinguistic variation limited to soft constraints

6.1.2.1. Causal Adequacy

An adequate model of gradience has to explain *why* grammaticality is a gradient, rather than a binary notion. Such an explanation is typically provided by stating a *source* for gradience in the grammar, i.e., the model has to specify which parts of the grammar are affected by gradience, and which parts are not.

6.1.2.2. Conceptual Adequacy

A model of gradience has to provide an intuitively correct concept of gradient grammaticality. This means that the notion of gradience provided by the model should be adequate for the measurement scale that is used to measure the data to be accounted for by the model. (See Section 2.3.1 for a more detailed discussion of measurement scales.)

Gradient data can be measured on an ordinal scale, such as the conventional scale "?", "??", "?*", "*" used in most of theoretical linguistics. Based on ordinal judgments of this sort, a model of gradience should provide a *qualitative* notion of gradience, i.e., the model should be able to make statements of the kind "structure S_1 is more grammatical than structure S_1".

Alternatively, gradient data can be measured on an interval scale. An interval scale provides measurements that quantify acceptability differences, instead of just specifying relative acceptability. This allows us to devise a *quantitative* notion of gradience, where the model supports statements of the sort "the grammaticality difference between structure S_1 and structure S_2 is 2.7".

6.1.2.3. Empirical Adequacy

To assess the empirical adequacy of a model, we have to take into account what type of data the model is based on. As argued in Chapter 2, experimentally collected acceptability judgments are the most appropriate source of data for models of gradient grammaticality. Some existing models of gradience, on the other hand, rely on intuitive judgment data, or corpus data.

Another criterion for empirical adequacy involves testing the model specifically

against the experimental data provided in Chapters 3 and 4. An empirically adequate model will be able to account for properties we summarized in Section 6.1.1, such as the cumulativity of constraint violations, context effects, and the soft/hard distinction.

6.1.2.4. Computational Adequacy

To assess the computational adequacy of a model, we ask if the model is available in a sufficiently precise algorithmic formulation. To meet this criterion, the model has to specify a *scoring algorithm*, i.e., an algorithm for computing the degree of grammaticality of a given structure based on the source of gradience the model assumes.

Secondly, the model has to specify a *training algorithm*, i.e., a way of estimating the model parameters from a set of data.[1] The data are typically gradient grammaticality judgments, while the parameters of the model depend on its assumptions about the source of gradience; constraint ranks or rule weights are typical examples.

An important property of the scoring algorithm and the parameter estimation algorithm is their *computational complexity*.

6.1.2.5. Predictive Adequacy

To assess the predictive adequacy of a model, we have to ask if the model provides a systematic way of evaluating the results of the scoring algorithm and the parameter estimation algorithm.

This question can be broken down into two subparts: *model fit* and *ability to generalize*. A model achieves a high model fit if it is able to account for the data it is trained on, i.e., after training, the model achieves a high match between the grammaticality scores it predicts and the acceptability scores it was trained on. Secondly, the model has to be able to generalize: it has to achieve good performance on unseen data, i.e., on data that has not been used for parameter estimation.

6.1.2.6. Cognitive Adequacy

Finally, we can discuss the cognitive plausibility of the model. This includes questions such as: (a) Are the linguistic representations that the model assumes cognitively plausible? (b) Can the predictions the model makes be related to findings about human language acquisition? (c) Does the model interact in a plausible way with models of human language processing?

[1] Note that the term training is used in a broad sense here. Of course models are possible that do not need training in the conventional sense; however, all models will require the adjustment of parameters so as to be able to account for a given set of data (in an OT-based framework the parameters are the constraint ranks).

6.2. Previous Models of Gradient Grammaticality

This section provides a survey of models of gradient grammaticality that have been proposed in the literature. We will distinguish three research traditions that developed models of gradience. The first one is in theoretical linguistics and will be discussed in Section 6.2.1. Gradience has also been the subject of some research in computational linguistics, which we will discuss in Section 6.2.2. Recently, some modeling efforts have been made in Optimality Theory; this is the subject of Section 6.2.3.

6.2.1. Theoretical Linguistics

6.2.1.1. Early Generative Grammar

The relevance of gradience for linguistic theory has been recognized early on in generative linguistics, with Chomsky (1955) probably being the first to discuss the issue. Chomsky (1955, 1964) proposes a model of gradient grammaticality that builds on a generalized notion of syntactic category. He assumes a hierarchy of orders (levels) of syntactic analysis, where each order is based on a set of categories that is more fine-grained than the set of categories of the previous order. In this approach, a sentence receives analyses of various orders, and it is grammatical of order n if it receives an analysis of order n on the category hierarchy. The highest-degree grammatical sentences are those with analyses of order one, corresponding to the most fine-grained set of categories. If a sentence only receives an analysis of order $n > 1$, its grammaticality is reduced, and the ungrammaticality increases with increasing n. (See Katz 1964 for a critical analysis of Chomsky 1964 and for an alternative proposal.)

This model of degrees of grammaticality is then refined by Chomsky (1965: Ch. 4). Here, the assumption of a hierarchy of syntactic categories is made explicit by decomposing categories into features and defining a partial order on the features. The phrase structure rules of a grammar then specify complex categories (bundles of features) instead of atomic nonterminal symbols. To model gradience, it is assumed that a phrase structure rule can be relaxed in that the feature specification of a lexical entry may deviate from the feature specification required by the rule. The structure derived by such a relaxed rule then is ungrammatical, where the ungrammaticality is greater the higher in the hierarchy is the feature whose specification is relaxed.

Chomsky's (1965) model of gradience has generated some experimental research (reviewed by Schütze (1996)), whose results are largely consistent with the predictions of the model, thus confirming its empirical adequacy. However, the model lacks computational and predictive adequacy, as no algorithms are available for estimating the model parameters (determining which rules to relax), and it is not obvious how the model's ability to generalize can be tested.

Note also that Chomsky's (1965) approach does not seem to carry over to more recent

generative frameworks. No explicit reference to gradience is made within Government and Binding Theory (Chomsky 1981) or the Minimalist Program (Chomsky 1995). This fact finds an explanation in the general tendency in contemporary generative grammar to develop more restrictive formalisms, and to attempt to integrate a notion of economy (a form of optimality) into generative models. Early generative grammar was descriptively rich, and a notion of gradience could be integrated fairly straightforwardly. This is no longer the case for current, more restrictive generative models.

6.2.1.2. Weighted Rules

Some researchers in the generative tradition have adopted weighted rule models as a way of accounting for degrees of grammaticality. An example is Uszkoreit's (1987) account of word order preferences. In this framework, grammatical rules are annotated with numeric weights that reflect their importance in determining grammaticality (for a similar proposal, see Jacobs 1988). Uszkoreit (1987) assumes constraint competition, i.e., not all constraints are necessarily satisfiable in a given linguistic structure. In this model, grammaticality is a gradient notion; the degree of grammaticality of a linguistic structure is computed as the sum of the weights of the constraint violations the structure incurs.

However, Uszkoreit's (1987) approach remains on an intuitive level; it is not founded on experimental evidence (but a partial experimental confirmation was later provided by Pechmann et al. 1994). Also, the Uszkoreit model only deals with word order variation; it remains to be seen if the approach is general enough to handle gradience in other parts of the grammar as well. Another problem is that Uszkoreit (1987) fails to make explicit how the rule weights are estimated from judgment data; it seems that this is left to the intuition of the linguist. This means that the model lacks computational adequacy. Also, no clear criteria are available to the assess model fit and the ability to generalize for an Uszkoreit-type model. Hence this approach also falls short of predictive adequacy.[2]

Another weighted grammar model is the Variable Rule approach proposed by Labov (1969) and Cedergren and Sankoff (1974) to account for sociolinguistic variation (mainly in phonology). This model differs from the one proposed by Uszkoreit (1987) in that it is probabilistic, and comes with an algorithm for parameter estimation (the Variable Rule model is essentially a log-linear model of the frequency distribution of grammatical forms in a corpus). Other weighted rule models have been proposed in the Fuzzy Grammar/Category Squish tradition (see Lakoff 1973; Mohan 1977; Mohanan 1993; Quirk 1965; Ross 1972, 1973a,b).

The Variable Rule model is in principle compatible with the data presented in this thesis; the ranking of constraints can be naturally expressed through rule weights. In this setting, hard constraints receive high weights, while soft ones receive low weights. Furthermore, the

[2] A model similar to Uszkoreit's (1987) has been proposed by Pafel (1998) to account for quantifier scope preferences. Essentially the same criticism applies.

Variable Rule model assumes that rule violations are cumulative, which corresponds to what we found in Chapters 3 and 4. It seems therefore possible to achieve empirical adequacy in a Variable rule framework.

The Variable Rule model also meets the criterion of computational adequacy, as it is based on quantitative data (corpus data), and is equipped with an estimation scheme that determines the rule weights from corpus frequencies. Nevertheless, it seems unlikely that the Variable Rule approach can be extended to deal with gradient phenomena like the ones investigated in this thesis. Gradient structures are typically extremely infrequent in a corpus (or fail to occur at all), and hence pose a serious sparse data problem for the estimation algorithm, making it difficult to obtain reliable estimates of the weights of gradient constraints. Note that this problem is an instance of the general unavailability of negative evidence from corpora—gradient structures constitute an instance of negative evidence (by virtue of being unacceptable, at least to a certain degree). (See also the general criticism of probabilistic grammar models in Section 6.2.2.2 below.)

Another more general objection concerns the conceptual adequacy of weight-based grammar models and is raised by Prince and Smolensky (1993). They formulate the problem in terms of a question that a potential opponent of OT might have:

> *Loss of Restrictiveness*. "In order to handle optimality, you must use numbers and use counting. The numerical functions required belong to a vast class which cannot be constrained in a reasonable way. Arbitrary quantization will be required, both in weighting degrees of concordance with (and violation of) individual constraints and in the weighting of the importance of disparate constraints with respect to each other. The result will be [a] system of complicated trade-offs (e.g. '*one* serious violation of $\mathbb{A}$ can be overcome when *three* moderate agreements with $\mathbb{B}$ co-occur with *two* excellent instances of $\mathbb{C}$.'), giving tremendous descriptive flexibility and no hope of principled explanation. Therefore, the main goal of generative grammatical investigation is irredeemably undermined." (Prince and Smolensky 1993: 197)

This question is answered affirmatively by Prince and Smolensky (1993) who contrast OT with weight based models, which suffer from a loss of restrictiveness:

> *Loss of Restrictiveness through Arithmetic.* Concern is well-founded here. As we have shown, however, recourse to the full-blown power of numerical optimization is not required. *Order*, not *quantity* (or *counting*), is the key in Harmony-based theories. In Optimality Theory, constraints are ranked, not weighted; harmonic evaluation involves the abstract algebra of order relations rather than numerical adjudication between quantities. (Prince and Smolensky 1993: 198)

Counterarguments against this view are presented by Guy (1997) and Guy and Boberg (1997), who show that cases of cumulative constraint violations cannot be efficiently represented in OT. Cumulativity, in their view, requires a probabilistic grammar model, such as the Variable Rule devised by Labov (1969) and Cedergren and Sankoff (1974) . (However, the Variable Rule approach has its problems when it comes to providing explanations for rule weights, as discussed by Anttila (1997).)

It is interesting in this context to consider another weighted rule model, viz., Harmonic Grammar (Legendre, Miyata, and Smolensky 1990a,b, 1991; Smolensky, Legendre, and Miyata 1992, 1993), a predecessor of OT that builds on the assumption that constraints are annotated with numeric weights (instead of just being rank-ordered as in Standard OT). Harmonic Grammar can be implemented in a hybrid connectionist-symbolic architecture and has been applied successfully to gradient data by Legendre et al. (1990a,b). As Prince and Smolensky (1993: 200) point out, "Optimality Theory [...] represents a very specialized kind of Harmonic Grammar, with exponential weighting of the constraints". We will provide a more detailed discussion of the relationship between Standard OT, Harmonic Grammar, and the model proposed in this thesis in Sections 6.4.3 and 6.4.4.

6.2.2. Computational Linguistics

6.2.2.1. Robust Parsing

Robust parsing is the task of assigning an analysis to a sentence even if the sentence is ungrammatical, e.g., because it contains errors such as typographical errors in the case of text, and slips of the tongue, disfluencies, or repairs in the case of speech. There is a large body of research in computational linguistics on robust parsing (for a survey, see the contributions in Carroll 1996). Robust parsing systems are typically not concerned with assigning a degree of grammaticality to a given ungrammatical sentence, but rather focus on how to recover from the ungrammaticality and provide an analysis even for sentences that are not generated by the grammar. An example for such a approach is the parsing scheme proposed by Core (1999) that deals with speech repairs in dialogues via meta-rules in the grammar.

However, there is a relevant research tradition in the computer aided instruction literature. For computer aided language instruction, robust parsing is typically utilized to deal with ungrammatical input provided by the student. In this setting, a system has to identify what type of ungrammaticality is present in the input, so as to provide adequate feedback to the student. This can be achieved, for instance, by selectively relaxing restrictions specified by the grammar (Kwasny and Sondheimer 1979; Robinson 1982; Weischedel and Black 1980).

As an example consider the relaxation approach proposed by Weischedel and Black (1980). In their system, grammar rules are associated with predicates that must be satisfied so that the grammar rule can be applied. These predicates are used to enforce, for instance,

subject-verb agreement. Weischedel and Black (1980) achieve robustness in their system by designating certain predicates as relaxable. For a given input, their system tries to return a parse that contains no relaxed predicates, but in the absence of such a parse, it returns the parse with the fewest relaxed predicates. Weischedel and Black (1980) show that this approach is useful for identifying why an input fails to parse and for providing explanatory feedback to the user.

Weischedel and Black's (1980) system is in effect based on a simple notion of degree of grammaticality, viz., the number of relaxed predicates that an input requires to parse successfully. An obvious problem with this approach is that this notion of degree of grammaticality is completely ad hoc: the grammar has to be hand-crafted, i.e., the grammar designer has to decide which predicates are relaxable. This means that the model falls short of computational adequacy. It shares this problem with the early generative models reviewed in Section 6.2.1.1; both approaches are based on rule relaxations.

6.2.2.2. Probabilistic Grammars

In the following, we will briefly discuss the relationship between gradience and probabilistic grammar models, based on the more detailed treatment of this topic by Keller (1996a: Ch. 3).

Probabilistic context-free grammars (PCFGs, see Manning and Schütze 1999 for an introduction) extend the formalism of context-free grammars (CFGs) by annotating each rule with a numeric value, viz., its probability. The annotations for all rules with the same left-hand side have to add up to one, and the probability of a parse is computed as the product of the probabilities of rules applied in that parse. For PCFGs, standard training methods like the inside-outside algorithm can be employed to train the rule probabilities from corpus data.

PCFGs constitute an efficient, well-understood technique for assigning probabilities to the analyses produced by a context-free grammar. They are commonly used for broad-coverage grammars, as CFGs large enough to parse unrestricted text typically are highly ambiguous, i.e., a single sentence will receive a large number of parses. This problem can be solved by training the grammar with a given corpus, where the training tries to maximize the overall probability of the corpus. The resulting probabilistic grammar then can be used to rank the analyses a sentence might receive, and improbable ones can be eliminated.

PCFGs are a straightforward probabilistic grammar formalism; a number of more sophisticated frameworks exist that enrich expressive linguistic formalisms with probabilistic information (Abney 1997; Brew 1994, 1995; Eisele 1994; Erbach 1993, 1997; Kim 1994; Riezler 1996, 1998). These frameworks share the property of assigning ranks to the analyses they produce. Therefore, it could be conjectured that they could be adapted so as to account for gradient grammaticality, with probabilities being reinterpreted as degrees of grammaticality.

However, as argued extensively by Keller (1996a: Ch. 3), the degree of grammaticality of a structure and its probability of occurrence in a corpus are two distinct concepts, and it

seems unlikely they can both be modeled in the same probabilistic framework. A related point of view is put forward by Abney (1996), who states that "[w]e must also distinguish degrees of grammaticality, and indeed, global goodness, from the probability of producing a sentence. Measures of goodness and probability are mathematically similar enhancements to algebraic grammars, but goodness alone does not determine probability. For example, for an infinite language, probability must ultimately decrease with length, though arbitrary long sentences may be perfectly good" (Abney 1996: 14). He also gives a number of examples for sentences that have very improbable, but perfectly grammatical readings. A similar point is made by Culy (1998), who argues that the statistical distribution of a construction does not bear on the question of whether it is grammatical or not.

Riezler (1996) agrees that probabilities and degrees of grammaticality are to be treated as separate concepts. He makes this point by arguing that, if one takes the notion of degree of grammaticality seriously for probabilistic grammars, there is no sensible application to the central problem of ambiguity resolution any more. A probabilistic grammar model cannot be trained so that the numeric value is assigned to a structure can function both as a well-formedness score (degree of grammaticality) and as a probability to be used for ambiguity resolution.

The right way of conceptualizing the difference between probability and gradience follows from the basic assumptions about competence and performance (see Section 2.2.2): disambiguation probabilities are part of linguistic performance, contributing to efficiency and robustness in language processing (ambiguity resolution, handling of distorted input), while gradience is part of a speaker's language competence.

Note that Keller and Asudeh (2000) present a similar argument in the context of Optimality Theory. They point out that if an OT grammar was to model both frequency and gradient grammaticality, then this would entail that the grammar incorporates both performance constraints (accounting for frequency effects) and competence constraints (accounting for grammaticality effects). This is highly undesirable in an OT setting, as it allows the crosslinguistic re-ranking of performance and competence constraints. Hence such a combined competence/performance grammar predicts that crosslinguistic differences can be caused by performance factors (e.g., memory limitations).

We conclude that probabilistic grammar models that fail to distinguish disambiguation probabilities and degrees of grammaticality and fall short of causal and conceptual adequacy. On the other hand, probabilistic grammars are computationally adequate as they typically come with an algorithm for parameter estimation (training algorithm). They also meet the criterion of predictive adequacy: a probabilistic model can be tested on unseen data to assess its ability to generalize.

6.2.3. Optimality Theory

In line with all major linguistic frameworks, Standard Optimality Theory (see Section 2.6 for an overview) assumes a binary notion of grammaticality: the competition between candidate structures selects one candidate (or a set of candidates sharing the same constraint profile) as optimal and, hence, grammatical. All losing candidates, i.e., those structures that are *suboptimal*, are assumed to be ungrammatical; Standard OT makes no predictions about the relative ungrammaticality of suboptimal candidates. This binary view of grammaticality is inadequate for data that exhibit a continuum of degrees of acceptability, such as the data reported in the present thesis. However, a number of proposals exist in the literature that propose to extend OT to deal with gradience; these will be reviewed in the present section.

6.2.3.1. Naive Extension of Standard Optimality Theory

A straightforward model of gradient grammaticality can be obtained extending the Standard OT notion of grammaticality. We will refer to this approach as the *Naive Extension of Standard Optimality Theory*.[3]

In Standard OT, grammaticality is defined as global optimality for the whole candidate set. This can be complemented by a definition of *suboptimality* as local optimality relative to a subset of the candidate set. We can then assume that a structure S_1 is less grammatical than a structure S_2 if S_1 is suboptimal with respect to S_2 (see Keller 1997 for a detailed proposal). Intuitively, this definition entails that the relative grammaticality of a structure corresponds to its *harmony*, i.e., its optimality theoretic rank in the candidate set. This model predicts a grammaticality ordering for the structures in the candidate set, which can then be tested against the acceptability ordering found experimentally for the candidates.

However, the Naive Extension of Standard OT encounters a number of serious problems when faced with experimental evidence such as the one presented in Chapters 3 and 4. One problem is that it predicts grammaticality differences *only* for structures in the same candidate set; relative grammaticality cannot be compared across candidate sets. The experimental findings, however, show that subjects can judge the relative grammaticality of arbitrary sentence pairs, a fact that cannot be accommodated by the Naive Extension of Standard OT.

Another problem is that grammaticality differences are predicted between *all* structures in a candidate set. A typical OT grammar assumes a richly structured constraint hierarchy, therefore all or most structures in a given candidate set will differ in optimality. The Naive Extension of Standard OT predicts that there is a grammaticality difference whenever there is a difference in optimality. This carries the danger that the Naive Extension of Standard OT will

[3]Throughout this chapter we will use this term to refer to a Naive Extension of Standard OT as proposed by Keller (1997) and discussed by Müller (1999). It is important to keep in mind that Standard OT was not designed to account for gradience, and hence our criticism applies only to this extension, not to Standard OT. We are not attacking Standard OT itself, as this would be attacking a straw man.

overgenerate, in the sense of predicting more grammaticality differences than are justified by the data (see Müller 1999 for more on this topic).

The cumulativity of constraint violations poses a third problem for the Naive Extension of Standard OT as a model of gradience. The experimental results demonstrate that the degree of ungrammaticality of a structure increases with the number of constraints it violates, both for soft and hard constraints. This fact is not accounted for by the Naive Extension: it relies on the Standard OT notion of optimality, which is defined via strict domination and predicts cumulativity effects only for constraints with the same ranking. Strict domination is incompatible with the cumulativity effect and the ganging up of constraint violations that was demonstrated in Chapters 3 and 4.

To summarize, the Naive Extension of Standard OT as a model of gradience has to be abandoned as it falls seriously short of empirical adequacy.

6.2.3.2. Floating Constraints

An alternative to the Naive Extension of Standard OT is the use of floating constraints, as proposed by Anttila (1997) and Nagy and Reynolds (1997). The core idea is to relax the ranking of certain constraints and allow them to "float" along the constraint hierarchy. This means that each candidate set is associated with more than one tableau, depending on how the ranking of the floating constraints is fixed. Anttila (1997) then assumes that the number of tableaux in which a given candidate is optimal predicts its probability of occurrence. He uses this approach to model the frequency distribution of a set of morphological forms in a corpus. A similar approach has been applied by Nagy and Reynolds (1997) for the modeling of corpus frequencies for phonological forms.

There are, however, a number of problems with the floating constraint approach, as pointed out by Boersma (1999b), Guy (1997), and Guy and Boberg (1997). The most serious one is that, just as the Naive Extension of Standard OT, a floating constraint approach is based on strict domination and therefore cannot account for the cumulativity effect and the ganging up of constraint violations attested experimentally. The floating constraint model therefore lacks empirical adequacy.

6.2.3.3. Grammaticality and Markedness

Müller (1999) discusses the shortcomings of the Naive Extension of Standard OT and proposes an alternative, the markedness model. This approach assumes a distinction between grammaticality (manifested in binary judgments) and markedness (associated with preferences). Grammaticality is handled in terms of Standard OT-style constraint competition. All candidates that are suboptimal in this competition are predicted to be categorically ungrammatical. For certain phenomena, the competition will produce not a single optimal candidate, but a set of optimal

candidates. All of these candidates are predicted to be grammatical; however, they take part in a further optimality theoretic competition based on a separate set of constraints, so-called markedness constraints. The optimal candidate in this competition is *unmarked*; the suboptimal candidates are more or less marked (dispreferred) depending on their relative suboptimality.[4]

Müller's (1999) model avoids the prediction of bogus grammaticality differences (a problem for the Naive Extension) as gradience is only induced by a subset of the constraints (the markedness constraints), which take part in a separate constraint competition. On the other hand, Müller's (1999) approach inherits from the Naive Extension of Standard OT the problem that only structures in the same candidate set can be compared as to their relative grammaticality.

Just as the Naive Extension and the floating constraint approach, Müller's (1999) model is unable to account for the cumulativity of constraint violations because the competition of markedness constraints uses the Standard OT constraint evaluation scheme based on strict domination. As a consequence, the empirical adequacy of the markedness approach is compromised.

6.2.3.4. Constraint Re-ranking

Keller (1998) suggests an alternative model of gradience that draws on concepts from OT learnability theory (Tesar and Smolensky 1998). As in the floating constraint model, constraint ranks assumed to be flexible and gradience is hypothesized to originate with this flexibility in the constraint hierarchy. The core idea of the re-ranking approach is to compute the relative grammaticality of a suboptimal structure by determining which constraint re-rankings are required to make the suboptimal structure optimal. This information can then be used to compare structures with respect to their degree of grammaticality: the assumption is that the degree of grammaticality of a candidate structure *S* depends on the number and type of re-rankings required to make *S* optimal. Such a re-ranking model offers the necessary flexibility to accommodate the experimental findings on constraint ranking and constraint interaction obtained in the experimental part of this thesis:

- The re-ranking model allows us to determine the relative grammaticality of arbitrary structures by comparing the number and type of re-rankings required to make them optimal. Comparisons of grammaticality are not confined to structures in the same candidate set, which accounts for the fact that subjects can judge the relative grammaticality of arbitrary sentence pairs.

- It seems plausible to assume that some constraint re-rankings are more serious than others, and hence cause a higher degree of ungrammaticality in the target structure.

[4]Note that Müller (1999) relies on the definition of markedness proposed by Höhle (1982: 102, 122): a given sentence S_1 is less marked than a sentence S_2 if it can occur in more context types than S_2. See Section 1.2.3 for a brief discussion.

This assumption allows us to model the experimental findings that some constraint violations are more serious than others. The experimental data justify two types of re-rankings, corresponding to the soft and hard constraint violations discussed above.

- The degree of grammaticality of a structure depends on the number of re-rankings necessary to make it optimal: the more re-rankings a structure requires, the more ungrammatical it becomes. This predicts the cumulativity of violations that was found experimentally both for soft and for hard constraints.

The re-ranking model offers a general way of dealing with degrees of grammaticality in OT, based on concepts that are independently motivated in OT learnability theory. However, a number of open questions remain.

An obvious problem concerns the cumulativity effect: if we assume that the degree of grammaticality of a given structure depends on the number of re-rankings it requires, then this naturally predicts that constraint violations are cumulative. However, this only holds for multiple violations of different constraints (requiring different re-rankings that are counted separately): multiple violations of the same constraint can be dealt with by a single re-ranking, and hence we fail to predict a cumulativity effect here. This prediction is not in accordance with the experimental facts: we found that the cumulativity of constraint violations extends to multiple violations of the same constraints (Experiment 6). A similar result was obtained by Chapman (1974), who investigated two types of violations (selectional restrictions and subcategorization requirements). Therefore, the empirical adequacy of the re-ranking model is compromised.

Another problem concerns the case of unmarked competitors. The algorithm proposed by Keller (1998) demotes the constraints violated by a structure S_1 below the ones violated by a given competitor S_2, so that S_1 becomes optimal. The degree of grammaticality of S_1 depends on the type and number of re-rankings required in this demotion process. Constraint demotion is impossible, however, if the competitor S_2 is completely unmarked, i.e., if it incurs no constraint violations at all, which entails that S_2 is optimal under any constraint ranking. The notion of relative grammaticality is not well-defined in this case, as no constraint demotion can take place. This means that the re-ranking model falls short of conceptual adequacy.

A third problem with the re-ranking model concerns computational adequacy, i.e., the absence of a training algorithm for determining constraint ranks from a set of judgment data: Keller (1998) does not provide a systematic method for computing an adequate constraint hierarchy based on a set of gradient linguistic judgments.

6.2.3.5. Probabilistic Optimality Theory and the Gradual Learning Algorithm

A further model of gradience in OT has been put forward by Boersma (1998, 2000) and Boersma and Hayes (2001).[5] This approach assumes a probabilistic variant of Optimality The-

[5]The present section owes a lot to discussions with Ash Asudeh and Paul Boersma.

ory and is designed to account for corpus frequencies and gradient acceptability judgments. It has been applied in phonology (Boersma 1997, 1998, 2000; Boersma and Hayes 2001; Boersma and Levelt 1999; Hayes 2000; Hayes and MacEachern 1998), morphology (Boersma and Hayes 2001; Hayes 1997b), and syntax (Asudeh 2001). We will refer to this model as *Probabilistic Optimality Theory* (POT).

The POT model stipulates that optimality-theoretic constraints are annotated with numeric weights; if a constraint C_1 has a higher weight that a constraint C_2, then C_1 outranks C_2. Boersma and Hayes (2001) assume *probabilistic constraint evaluation*, which means that at evaluation time, a small amount of random noise is added to the weight of a constraint. As a consequence, re-rankings of constraints are possible if the amount of noise added to the weights exceeds the difference between the weights of the constraints.

For instance, assume that two constraints C_1 and C_2 are ranked $C_1 \gg C_2$, selecting the structure S_1 as optimal for a given input. In POT, a re-ranking of C_1 and C_2 can occur at evaluation time, resulting in the opposite ranking $C_2 \gg C_1$. This re-ranking might result in an alternative optimal candidate S_2. The probability of the re-ranking that makes S_2 optimal depends on the difference between the weights of C_1 and C_2 (and on the amount of noise added to the weights). The re-ranking probability is assumed to predict the corpus frequency of S_2. The more probable the re-ranking $C_2 \gg C_1$, the higher the corpus frequency of S_2; if the rankings $C_1 \gg C_2$ and $C_2 \gg C_1$ are equally probable, then S_1 and S_2 have the same corpus frequency. Boersma and Hayes (2001) assume that corpus frequency and degree of grammaticality are directly related, which means that POT also provides a model of gradient grammaticality.

The POT model comes with its own learning theory in the form of the Gradual Learning Algorithm (GLA; Boersma 1998, 2000; Boersma and Hayes 2001). This algorithm is a generalization of Tesar and Smolensky's (1998) Constraint Demotion Algorithm: it performs constraint promotion as well as demotion.

More specifically, the GLA works as follows. It starts with a grammar G, in which initially the constraints are ranked arbitrarily, i.e., they have random weights. If the GLA encounters a training example S, it will compute the corresponding structure S' currently generated by the grammar G. If S and S' are not identical, then learning takes place; the constraint hierarchy of G has to be adjusted such that it makes S optimal, instead of S'. (Note that S is a training example and thus known to be grammatical.) In order to achieve this, the GLA performs the following steps: (a) it decreases (by a small amount) the weights of all constraints that are violated by S but not by S'; (b) it increases (by a small amount) the weights of all constraints that are violated by S' but not by S. This procedure will gradually adjust the weights of the constraints in G, resulting ultimately in the correct constraint hierarchy (given that enough training data is available).

In contrast to the re-ranking model, POT is computationally adequate as it is equipped with a training algorithm (viz., the GLA outlined above). Boersma and Hayes (2001) test the

Table 6.2: Data set that cannot be captured by POT (hypothetical acceptability scores)

	C_3	C_1	C_2	Acceptability
S_1		*		3
S_2		*	*	2
S_3	*			1

algorithm on certain data sets from morphology and phonology and demonstrate that the algorithm achieves a good model fit. However, no testing on unseen data is carried out, hence it remains unclear if the GLA is able to generalize, rather than just fitting the training data.[6] Also, no proof of the correctness is available for the GLA, and its convergence properties are unknown (although a sketch of a correctness proof is provided in Boersma 1998).

Just as the re-ranking model, POT has the advantage of allowing us to compare the relative grammaticality of arbitrary structures. The seriousness of constraint violations is readily accounted for by the probability of the re-ranking required to make the suboptimal candidate optimal. Also, the cumulativity effect can be modeled by assuming that the ungrammaticality caused by multiple constraint re-rankings is cumulative.

However, there seem to be problems with the empirical adequacy of POT. Like the re-ranking approach (see Section 6.2.3.4), the POT approach is not able to deal with unmarked competitors—an unmarked competitor does not incur any constraint violations, i.e., it is always optimal, no matter which re-rankings are assumed. Furthermore, it seems that certain examples involving cumulative constraint violations cannot be captured by POT. In the following, we will discuss this point in more detail.

First, we provide an example that illustrates that there seem to be data sets that cannot be modeled in POT. Assume two structures S_1 and S_2 in the same candidate set, which both incur a violation of the constraint C_1. The structure S_2 incurs an additional violation of the constraint C_2, and S_1 and S_2 incur no other violations (or incur the same violations). Now assume a third structure S_3 that only incurs a violation of the constraint C_3. Assume further that S_2 is less grammatical than S_1 and let S_3 be less grammatical than S_2.

This configuration is illustrated in Table 6.2. POT does not seem to be able to model this data set: there is no re-ranking under which S_2 is optimal, as S_2 incurs the same violations as S_1, plus an additional violation of C_2. Hence S_1 will always win over S_2, no matter which constraint re-rankings we assume. Under the POT approach, the degree of grammaticality of a structure depends on how likely it is for this structure to be optimal. S_2 can never be optimal, it is a "perpetual loser" and therefore is predicted to be categorically ungrammatical. S_3, on the other hand, is not a perpetual loser, as there are re-rankings which make it optimal (e.g., $C_1 \gg C_3$ and

[6] However, Paul Boersma (personal communication, December 2000) reports that he has carried out extensive tests on unseen data using the data sets for Ilokano and Finnish presented by Boersma and Hayes (2001). These tests show that the GLA achieves a good model fit on the test data, i.e., that it is able to generalize.

Table 6.3: Data set that cannot be captured by POT, taken from Experiment 10 (log-transformed mean acceptability scores)

{V, S, O}	VERBFIN	NOMAGN	PROAGN	Acceptability
$O_{pro}SV$		*		.2412
$OS_{pro}V$		*	*	−.0887
$VS_{pro}O$	*			−.1861

$C_2 \gg C_3$). This means that a situation where S_3 is less grammatical than S_2 cannot be captured by a POT grammar.

Configurations such as this one occur in the experimental data presented in Chapters 4 and 3; they are not just artificially constructed counterexamples. Consider an example from Experiment 10. Recall that we assumed three constraints: VERBFINAL specifies that the verb has to be in final position, NOMALIGN specifies that nominative NPs have to precede accusative NPs, while PROALIGN states that pronouns have to precede full NPs. Table 6.3 lists the experimental acceptability scores for structures that incur one violation of NOMALIGN, one violation of VERBFINAL, and a combined violation of NOMALIGN and PROALIGN. The relative acceptability values match the ones in the example in Table 6.2, hence this examples poses a problem for POT.

The GLA (Boersma 1999a) is designed to learn POT grammars from acceptability data or frequency data. POT appears to be unable to capture configurations like the ones in Tables 6.2 and 6.3. This implies that the GLA should not be able to learn these configuration, a prediction that can be verified empirically using Praat (Boersma 1999a), a software package that implements the GLA. When confronted with a training set that contains the configuration in Table 6.2, the GLA fails to converge; a continuous downdrift of the weights of C_1, C_2, and C_3 is observed.

There is a related type of data that seems to pose problems for the POT approach. This concerns data sets containing cumulative violations of the same constraint. As an example, consider the data set in Table 6.4, where the winning candidate is S_1, incurring a violation of C_2. If a re-ranking $C_2 \gg C_1$ occurs, then S_2, incurring a single violation of C_1, will win. However, there is no re-ranking that can make S_3 or S_4 optimal, as these candidates have the same violation profile as S_2, but incur multiple violations of C_1. This means that POT predicts that there should be no cumulative effects from multiple constraint violations: all structures that incur n violations of a given constraint ($n > 1$) will be equally ungrammatical (provided they are minimal pairs, i.e., they share the same constraint profile on all other constraints). This prediction is at odds with results that demonstrate cumulativity effects for multiple violations of the same constraint (see Experiment 6).

Paul Boersma (personal communication, December 2000) draws attention to the fact

Table 6.4: Example of a data set with cumulative effects

	C_1	C_2
S_1		*
S_2	*	
S_3	**	
S_4	***	

that he is aware of several other apparent counterexamples to the POT model, including certain data sets presented by Reynolds (1994). He points out that a reanalysis with a different constraint set makes it possible for POT to model these data. This leads him to conjecture that also the counterexamples presented in this thesis can be reanalyzed with a different set of constraints, enabling POT to capture these data sets, and the GLA to learn them. We will leave this question for future research.

6.3. Linear Optimality Theory

The aim of this section is to propose a grammar model that accounts for the properties of gradient data discussed in Section 6.1.1 and meets the criteria for models of gradience outlined in Section 6.1.2, while avoiding the pitfalls of earlier attempts to model gradience in grammar that were discussed in Section 6.2.

In order to satisfy these desiderata, we propose a model of gradience that makes predictions about the relative grammaticality of linguistic structures. Our model builds on core concepts from Optimality Theory, a framework that is attractive for our purposes as it is equipped with a notion of competition that allows us to formalize the interaction of linguistic constraints. Furthermore, OT provides a notion of constraint ranking that allows us to account for the fact that constraints differ in strength, i.e., that some constraints are more important than others for the overall well-formedness of a given linguistic structure.

Although the model we propose borrows central concepts (such as constraint ranking and competition) from Optimality Theory, it differs in two crucial respects from existing OT-based accounts. Firstly, we assume that constraint ranks are represented as sets of numeric weights, instead of as partial orders. Secondly, we assume that the grammaticality of a given structure is proportional to the sum of the weights of the constraints it violates. This means that we replace OT's notion of strict domination with an linear constraint combination scheme. We will therefore call the model we propose Linear Optimality Theory (LOT).

In Section 6.3.7, it will be argued that the linear combination scheme is well supported by the data from Experiments 1–12. Furthermore, it will be shown that the other properties of gradient data (soft/hard distinction, context effects, crosslinguistic effects) follow from the

linearity assumption. The LOT approach has the added advantage of permitting the use of standard model fitting algorithms, such as Least Square Estimation, to compute the constraint weights.

6.3.1. Components of an OT Grammar

Only a limited number of components of the OT architecture are affected by the proposals we make. The changes concern only *HEval*, the function that evaluates the harmony of a candidate, and *Rank*, the ranking component. Our proposal does not affect assumptions concerning the input and the generation function *Gen*, the two components of an OT grammar that determine which structures compete with each other. Also the constraint component *Con*, i.e., formal apparatus for representing constraints and candidates is unaffected. Our proposals are neutral in these respects, and compatible with the diverse assumptions put forward in the OT literature.

However, our new version of *HEval* and *Rank* entail changes in the way the optimal candidate is computed, as well as requiring a new type of ranking argumentation, i.e., a method for establishing constraint ranks from a set of linguistic examples. It will be shown that this type of ranking argumentation is considerably simpler than the one classically assumed in OT. Also, well understood algorithms exist for automating this type of ranking argumentation.

6.3.2. Violation Profiles and Harmony

The most prominent pattern in the experimental data presented in Chapters 3 and 4 was the *cumulativity* of constraint violations, i.e., the fact that the degree of unacceptability of a structure increases with the number of constraint violations it incurs. Cumulativity was in evidence in the extraction data presented in Experiment 4, in the binding data in Experiment 5, in the gapping data in Experiment 8, and in the word order data in Experiments 6 and 10. It was shown that both soft and hard constraint violations are cumulative (see Experiment 4), and that the cumulativity effect extends from multiple violations of different constraints to multiple violations of the same constraint (see Experiment 6).

The other pervasive pattern in the data was the *ranking* of constraints, i.e., the fact that constraint violations differ in the degree of unacceptability they cause. Constraint ranking was observed in the extraction data presented in Experiments 4 and 9, in the binding data in Experiment 5, in the gapping data in Experiment 8, and in the word order data in Experiments 6 and 10–12. Again, the ranking of constraints seems to hold for both soft constraints (see Experiments 8, 6, 10–12) and for hard constraints (see Experiments 4 and 9).

The model of gradient grammaticality that this thesis advocates derives from these two fundamental findings about constraint cumulativity and constraint ranking. We will adopt two hypotheses to implement these two results. The first hypothesis deals with constraint ranking:

(6.1) **Ranking Hypothesis**

The ranking of linguistic constraints can be implemented by annotating each constraint with a numeric weight representing the reduction in acceptability caused by a violation of this constraint.

In Section 3.1.2, we put forward an operational definition of constraint ranking as the degree of unacceptability caused by a constraint violation, and in Chapters 3 and 4 we showed that this definition is a useful tool in experimentally assessing the strength of constraint violations. The Ranking Hypothesis in (6.1) goes one step further: it assumes that constraint rankings can be modeled in terms of numeric weights representing the reduction in acceptability caused by constraint violations.

Note that this notion of constraint ranks as numeric weights is more general than the notion of ranks standardly assumed in Optimality Theory. Standard OT formulates constraint ranks as binary ordering statements of the form $C_1 \gg C_2$, meaning that constraint C_1 is ranked higher than the constraint C_2. Such statements do not make any assumptions regarding *how much* higher the ranking of C_1 is compared to the ranking of C_2 (but see Boersma 1998 for a version of OT where the constraint ranks are quantified numerically). Such information is only available once we adopt a numeric concept of constraint ranking.

In the remainder of the thesis, we will adopt the following terminological convention. The term constraint *weight* will be used to refer to the numeric annotation that our model assigns to a constraint. The term constraint *rank* will be employed to refer to the relative weight of two constraints in our model: we say that a constraint outranks another constraint if it has a greater weight (see also Definition (6.9) below). Thirdly, we will retain the use of the term rank to refer to a constraint ordering. This usage is justified by the fact that Standard OT ranks (i.e., constraint orderings) are a special case of ranks as defined in Linear Optimality Theory (this will be shown in Section 6.3.8).

Once numeric constraint ranks have been postulated, the overall acceptability of a structure can be computed based on the constraints that the structure violates. We will assume that simple summation is sufficient to compute the degree of acceptability of a structure from the weights of the constraints that the structure violates. This will account straightforwardly for the cumulativity of constraint violations that was in evidence in the experimental data throughout this thesis. In Chapter 7 we will demonstrate that this approach achieves a good model fit on data sets based on the experimental results from Chapters 3 and 4.

To account for the cumulativity of constraint weights, we formulate the Linearity Hypothesis in (6.2), which is at the core of Linear Optimality Theory, the model of gradience we propose.

(6.2) **Linearity Hypothesis**

The cumulativity of constraint violations can be implemented by assuming that the

grammaticality of a structure is proportional to the weighted sum of the constraint violations it incurs, where the weights correspond to constraint ranks.

To make explicit the hypotheses in (6.1) and (6.2), we will formulate a numeric model that relates constraints ranks and degree of grammaticality. We first define the notion of a grammar signature, which specifies the constraint set and the associated weights for a grammar. (Note that this definition, and all subsequent ones, are independent of the formulation of the constraints proper; our account is one of constraint interaction, not of actual linguistic constraints.)

(6.3) **Grammar Signature**

A grammar signature is a tuple $\langle \mathbf{C}, w \rangle$ where $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$ is the constraint set, and $w(C_i)$ is a function that maps a constraint $C_i \in \mathbf{C}$ on its constraint weight w_i.

Relative to a grammar signature, a given candidate structure has a constraint violation profile as defined in (6.4). The violation specifies which constraints are violated by the structure and how often. This is a useful auxiliary notion that we will rely on in further definitions.

(6.4) **Violation Profile**

Given a constraint set $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$, the violation profile of a candidate structure S is the function $v(S, C_i)$ that maps S on the number of violations of the constraint $C_i \in \mathbf{C}$ incurred by S.

Based on Definitions (6.3) and (6.4), we can now define the harmony of a structure using a simple linear model:

(6.5) **Harmony**

Let $\langle \mathbf{C}, w \rangle$ be a grammar signature. Then the harmony $H(S)$ of a candidate structure S with a violation profile $v(S, C_i)$ is given in (6.6).

(6.6) $$H(S) = -\sum_i w(C_i) v(S, C_i)$$

Equation (6.6) states that the harmony of a structure is the negation of the weighted sum of the constraint violations that the structure incurs. Intuitively, the harmony of a structure describes its degree of well-formedness relative to a given set of constraints. This notion corresponds closely to the definition of harmony assumed in Standard OT (Prince and Smolensky 1997: 1607) or Harmonic Grammar (Smolensky et al. 1992: 14).

We will assume that all constraint weights are positive, i.e., that $w_i \geq 0$ for all i. This means that only constraint violations influence the harmony of a structure. Constraint satisfactions will not change the harmony of the structure (including cases where a constraint is vacuously satisfied because it is not applicable). This assumption is in accordance with the experimental results in Chapters 3 and 4, where only constraints violations were found to affect acceptability.

In the next section, we will see that harmony is related to grammaticality in a way that implements the Linearity Hypothesis in (6.2).

6.3.3. Constraint Competition and Optimality

Based on the definitions of violation profile and harmony proposed in the preceding section, we can now specify a notion of grammaticality in Linear Optimality Theory. We define grammaticality in terms of the relative harmony of two candidates in the same candidate set:

(6.7) **Grammaticality**
Let S_1 and S_2 be candidate structures in the candidate set **R**. Then S_1 is more grammatical than S_2 if $H(S_1) > H(S_2)$. This can be abbreviated as $S_1 > S_2$.[7]

A crucial difference between harmony and grammaticality follows from Definition (6.7). Harmony is an absolute notion that describes the overall well-formedness of a structure. Grammaticality, on the other hand, describes the relative ill-formedness of a structure compared with another structure. While it is possible to compare the harmony of two structures across candidates sets, the notion of grammaticality is only well-defined for two structures that belong to the same candidate set (i.e., share the same input). Therefore, Definition (6.7) (and the subsequent Definition (6.8)) provide a *relative* notion of well-formedness, in line with the optimality theoretic tradition.

Based on the definition of grammaticality in (6.7), we can define the optimal structure in a candidate set as the one with the highest relative grammaticality.

(6.8) **Optimality**
A structure S_{opt} is optimal in a candidate set **R** if $S_{opt} > S$ for every $S \in \mathbf{R}$.

A notion of constraint rank can readily be defined in LOT based on the relative weight of two constraints (see also the terminological note on ranks vs. weights in Section 6.3.2 above):

(6.9) **Constraint Rank**
A constraint C_1 outranks a constraint C_2 if $w(C_1) > w(C_2)$. This can be abbreviated as $C_1 \gg C_2$.

In what follows, we will illustrate the definitions for harmony, grammaticality, and optimality. Consider an example grammar with the constraints C_1, C_2, and C_3, and the constraints weights given in Table 6.5. This table also specifies an example candidate set $S_1, \ldots, S_4$ and gives the violation profiles for these candidates. The harmony for each of these structures can be computed based on Definition (6.5).

The structure S_3 maximizes harmony, i.e., it incurs the least serious violation profile. It is therefore the optimal structure in the candidate set, i.e., it is more grammatical than all other candidate structures. The structures S_1 and S_4 are both less grammatical than S_3. S_1 and

[7] This usage differs from the standard OT usage, where harmonic ordering is denoted by "$\succ$", not "$>$". The symbol "$\succ$" is already used for constituent ordering in this thesis (see Section 3.7.1).

Table 6.5: Example violation profile and harmony scores

	C_1	C_2	C_3	
$w(C)$	4	3	1	$H(S)$
S_1		*	*	-4
S_2		*	**	-5
S_3			*	-1
S_4	*			-4

S_4 receive the same harmony scores, but for different reasons; S_4 because it incurs a high-ranked violation of C_1, S_1 because it accumulates violations of C_2 and C_3. The structure S_2 is less grammatical than S_1, as it incurs an additional violations of C_3. In total, we obtain the following grammaticality hierarchy: $S_3 > \{S_1, S_4\} > S_2$.

This examples illustrates the three central properties of constraint interaction that were identified in Chapters 3 and 4. The first property is the *ranking* of constraints. S_3 incurs a violation of C_3, while S_4 incurs a violation of C_1. That S_3 is more grammatical than S_4 is accounted for by the fact that C_1 has a higher weight than C_3, i.e., the ranking $C_1 \gg C_3$ holds. This is a situation that was observed many times in the experimental data presented in Chapters 3 and 4.

Furthermore, the example illustrates how the *cumulativity* of constraint violations is modeled. S_1 incurs single violations of C_2 and C_3. The structure S_2 also incurs a single violation of C_2, but a double violation of C_3. As a consequence, S_1 is more grammatical than S_2. Cumulativity effects such as these encountered frequently in the experimental data in Chapters 3 and 4.

Finally, Table 6.5 illustrates the *ganging up* of constraint violations. The structures S_1 and S_4 have different constraint profiles: S_4 violates the constraint C_1, while S_1 violates the two constraints C_2 and C_3, which are both lower ranked than C_1. However, S_1 and S_4 are equally grammatical because the two constraints C_2 and C_3 gang up against C_1, leading to the same harmony score in both structures. Ganging up effects were encountered in Experiments 4, 5, and 10.

Note that standard optimality theoretic evaluation of the candidate set in Table 6.5 leads to a different harmonic ordering: $S_3 > S_1 > S_2 > S_4$. Under the Naive Extension of Standard OT (see Section 6.2.3.1), this order corresponds to the grammaticality order of the candidates. The Naive Extension assumes the strict domination of constraints, and therefore fails to model ganging up effects. Under this approach, there is no possibility for a joint violation of C_2 and C_3 to be as serious as a single violation of C_1, due to the ranking $C_1 \gg C_2 \gg C_3$. Hence the Naive Extension of Standard OT fails to account for the ganging up effects that were observed experimentally.

The example in Table 6.5 also demonstrates a problem with re-ranking approaches to gradience, as discussed in Sections 6.2.3.4 and 6.2.3.5. Structure S_2 is less grammatical than

structure S_1, as it incurs a double violation of C_3, while S_1 only incurs a single violation. However, S_2 is more grammatical than the competitor S_4, which incurs a violation of the more highly ranked constraints C_1. This situation cannot be modeled by the re-ranking model. There is no re-ranking under which S_2 can be optimal: it will alway incur one more violation than S_1, no matter how C_2 and C_3 are ranked. The candidate S_2 is a "perpetual loser" and is predicted to be categorically ungrammatical. In particular, this means that S_2 is predicted to be less grammatical than S_4, which can become grammatical by virtue of a re-ranking of C_1. (See Section 6.2.3.5 for a more detailed explanation of the problem of perpetual losers.)

6.3.4. Ranking Argumentation

Optimality Theory employs so-called *ranking arguments* to establish constraint rankings from data. A ranking argument refers to a set of candidate structures with a certain constraint violation profile, and derives a constraint ranking from this profile. This can be illustrated by the following example: assume that two structures S_1 and S_2 have the same constraint profile, with the following exception: S_1 violates constraint C_1, but satisfies C_2. Structure S_2, on the other hand, violates constraint C_2, but satisfies C_1. If S_1 is acceptable but S_2 is unacceptable, then we can conclude that the ranking $C_2 \gg C_1$ holds (see Prince and Smolensky 1993: 106).

In the general case, the fact that S_1 is acceptable but S_2 is unacceptable entails that each constraint violated by S_1 is outranked by at least one constraint violated by S_2. (See Hayes 1997a for a more extensive discussion of the inference patterns involved in ranking argumentation in Standard OT.)

The LOT approach advocated in this thesis allows a form of ranking argumentation that relies on gradient acceptability data instead of the binary acceptability judgments used in Standard OT. A ranking argument in Linear Optimality Theory can be constructed based on the difference in acceptability between two structures in the same candidate set, using the following definition:

(6.10) **Ranking Argument**
Let S_1 and S_2 be candidate structures in the candidate set **R** with the acceptability difference ΔH. Then the equation in (6.11) holds.

(6.11) $H(S_1) - H(S_2) = \Delta H$

This definition assumes that the difference in harmony between S_1 and S_2 is accounted for by ΔH, the acceptability difference between the two structures. ΔH can be observed empirically, and measured, for instance, using magnitude estimation judgments such as the ones collected in Chapter 3 and 4. Drawing on the definition of harmony in (6.5), Equation (6.11) can be transformed to:

(6.12) $\sum_i w(C_i)(v(S_1,C_i) - v(S_2,C_i)) = -\Delta H$

This assumes that S_1 and S_2 have the violation profiles $v(S_1)$ and $v(S_2)$ and are evaluated relative to the grammar signature $\langle \mathbf{C}, w \rangle$.

Typically, a single ranking argument is not enough to rank the constraints of a given grammar. Rather, we need to accumulate a sufficiently large set of ranking arguments, based on which we can then deduce the constraint hierarchy of the grammar. To obtain a maximally informative set of ranking arguments, we take all the candidate structures in a given candidate set and compute a ranking argument for each pair of candidates, using Definition (6.12).

The number of ranking arguments that a set of k candidates yields is given in (6.13); note that this is simply the number of all unordered pairs that can be generated from a set of k elements.

$$n = \frac{k^2 - k}{2} \tag{6.13}$$

Now we are faced with the task of computing the constraint weights of a grammar from a set of ranking arguments. This problem can be solved by regarding the set of ranking arguments as a system of linear equations. The solution for this system of equations will then provide a set of constraint weights for the grammar. This idea is best illustrated using an example. We consider the candidate set in Table 6.5 and determine all ranking arguments generated by this candidate set (here w_i is used as a shorthand for $w(C_i)$, the weight of constraint C_i):

$$\begin{array}{lrcrcr}
S_1 - S_2: & 0w_1 + 1w_2 + 1w_3 - 0w_1 - 1w_2 - 2w_3 & = & -((-4) - (-5)) & = & -1 \\
S_1 - S_3: & 0w_1 + 1w_2 + 1w_3 - 0w_1 - 0w_2 - 1w_3 & = & -((-4) - (-1)) & = & 3 \\
S_1 - S_4: & 0w_1 + 1w_2 + 1w_3 - 1w_1 - 0w_2 - 0w_3 & = & -((-4) - (-4)) & = & 0 \\
S_2 - S_3: & 0w_1 + 1w_2 + 2w_3 - 0w_1 - 0w_2 - 1w_3 & = & -((-5) - (-1)) & = & 4 \\
S_2 - S_4: & 0w_1 + 1w_2 + 2w_3 - 1w_1 - 0w_2 - 0w_3 & = & -((-5) - (-4)) & = & 1 \\
S_3 - S_4: & 0w_1 + 0w_2 + 1w_3 - 1w_1 - 0w_2 - 0w_3 & = & -((-1) - (-4)) & = & -3
\end{array} \tag{6.14}$$

This system of linear equations can be simplified to:

$$\begin{array}{lcr}
-w_3 & = & -1 \\
w_2 & = & 3 \\
w_2 + w_3 - w_1 & = & 0 \\
w_2 + w_3 & = & 4 \\
w_2 + 2w_3 - w_1 & = & 1 \\
w_3 - w_1 & = & -3
\end{array} \tag{6.15}$$

We have therefore determined that $w_2 = 3$ and $w_3 = 1$. The value of w_1 can be easily be obtained from any of the remaining equations: $w_1 = w_2 + w_3 = 4$.

This example demonstrated how a system of linear equations that follows from a set of ranking arguments can be solved by hand. However, such a manual approach is not practical for large systems of equations as they occur in realistic ranking argumentation. Typically, we

will be faced with a large set of ranking arguments, generated by a candidate set with many structures, or by several candidate sets.

There are a number of standard algorithms for solving systems of linear equations, which can be utilized for automatically determining the constraint weights from a set of ranking arguments. In the next section, we will discuss Gaussian Elimination, an algorithm which delivers an exact solution of a system of linear equations (if there is one), and Least Square Estimation, which determines an approximate solution. We will argue that the Least Square Estimation is well-suited for estimating constraints weights from experimental data such as the one presented in Chapters 3 and 4.

6.3.5. Algorithms for Estimating Constraint Ranks

In this section, we will deal with the problem of parameter estimation, which for Linear Optimality Theory amounts to the task of determining the constraint weights of a grammar from a set of ranking arguments. In the previous section, we showed that this task can be reduced to solving a system of linear equations, a well-understood mathematical problem. In what follows, two algorithms will be described that solve this problem, Gaussian Elimination and Least Square Estimation.

6.3.5.1. Gaussian Elimination

A system of linear equations can be represented as an augmented matrix. In this representation, the coefficients of the equations correspond to the matrix elements, while a separate column on the right contains the constants on the right hand sides of the equations. As an example, consider the following augmented matrix representing the system of equations in (6.14):

(6.16)
$$\left[\begin{array}{rrr|r} 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 3 \\ -1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 4 \\ -1 & 1 & 2 & 1 \\ -1 & 0 & 1 & -3 \end{array}\right]$$

By transforming the augmented matrix into *echelon form*, we can solve the corresponding system of linear equations. This transformation can be achieved by performing *Gaussian Elimination* on the matrix. The algorithm for Gaussian Elimination is given in Figure 6.1

To illustrate how this algorithm works, we will solve the system of linear equations in (6.14), represented as an augmented matrix in (6.16). We start with the element in the first

1. Let i and j be a counter for the row and the column of the matrix, respectively. Let $i = 1$ and $j = 1$.
2. If element$(i, j) = 0$ then swap row i with a row below row i so that element$(i, j) \neq 0$. if there is no such row, increment j and repeat step 2.
3. Divide all elements of row i by element(i, j), so that element$(i, j) = 1$.
4. For each row below row i obtain a zero in column j by subtracting a multiple of row i from that row.
5. Stop if i is equal to the maximum number of rows, else increment i and j and go to step 2.

Figure 6.1: Algorithm for Gaussian Elimination

column of row 1. Because this element is 0, we have to swap row 1 and row 3 and obtain:

$$\left[\begin{array}{rrr|r} -1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & -1 & -1 \\ 0 & 1 & 1 & 4 \\ -1 & 1 & 2 & 1 \\ -1 & 0 & 1 & -3 \end{array}\right] \tag{6.17}$$

Now we multiply all elements of row 1 with -1 to obtain a 1 in the first column of row 1:

$$\left[\begin{array}{rrr|r} 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & -1 & -1 \\ 0 & 1 & 1 & 4 \\ -1 & 1 & 2 & 1 \\ -1 & 0 & 1 & -3 \end{array}\right] \tag{6.18}$$

The next step is to zero all elements in the first column below row 1 by subtracting a multiple of row 1. Only rows 5 and 6 need to be zeroed, with a multiplication factor of -1 in both cases.

This yields the following matrix:

$$\left[\begin{array}{ccc|c} 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & -1 & -1 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 1 & 1 \\ 0 & -1 & 0 & -3 \end{array}\right] \tag{6.19}$$

Now we move on to row 2. Here, the element in column 2 is already 1. We just have to zero rows 4 and 6 by subtracting row 2:

$$\left[\begin{array}{ccc|c} 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right] \tag{6.20}$$

Now we proceed to row 3, which we first multiply by -1 to obtain a 1 in the third column of row 3:

$$\left[\begin{array}{ccc|c} 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right] \tag{6.21}$$

Then rows 4 and 5 have to be zeroed, which yields:

$$\left[\begin{array}{ccc|c} 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \tag{6.22}$$

All remaining rows contain zeros, so the algorithm terminates; the resulting matrix is in echelon form. From this matrix, we can now obtain the solution of the system of equations by *backward substitution*. We work our way up from the last row of the matrix, skipping rows that only contain zeros. Row 3 then gives us the solution for w_3, viz., $w_3 = 1$. The next step is to consider row 2, which gives us $w_2 = 3$. Finally, row 1 represents the equation $w_1 - w_2 - w_3 = 3$, from which we can obtain w_1 by substituting the values for w_2 and w_3, which yields $w_1 = 4$.

A system of linear equations can be *consistent*, i.e., have one or more solutions, or *inconsistent*, i.e., have no solution. An example for an augmented matrix of a inconsistent system solution is:

$$\left[\begin{array}{ccc|c} 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \tag{6.23}$$

Here, the fourth column indicates a contradiction, it represents the equation $0 = 1$. This indicates that the system is inconsistent. In an LOT setting, this means that there is no set of constraint weights that fits the set of ranking arguments on which the system of equations was based.

A consistent system of linear equations can have exactly one solution (such as in our example above), or infinitely many solution. In the latter case, the echelon matrix is such that no fixed value can be assigned for one or more of the variables. As an example consider the following matrix:

$$\left[\begin{array}{ccc|c} 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \tag{6.24}$$

For this matrix, backward substitution gives us $w_2 = 3$ and $w_1 = w_3 + 3$. Infinitely many values of w_1 and w_3 satisfy these conditions. In an LOT setting, such a situation indicates that more than one set of weights is compatible with the set of ranking arguments on which the set of equations is based. This might mean that we need to consider more data (more or larger candidate sets) to obtain a unique set of weights of the grammar under consideration.

6.3.5.2. Least Square Estimation

Least Square Estimation (LSE) is a method for finding a solution for a system of linear equations even if the system is inconsistent. This means that LSE enables us to estimate the constraint weights of an LOT grammar if there is no set of weights that satisfy all the ranking arguments exactly (in contrast to Gaussian elimination). Rather, LSE will find an approximate set of constraint weights that maximizes the fit between with the acceptability scores. This is the right strategy for modeling experimental data such as the one we dealt with in Chapters 3 and 4.

In what follows, we will explain Least Square Estimation for the simple case of an equation system with only one variable. We opt for this simplification as it allows us to derive the weight equation in a straightforward manner. An equation system with one variable is represented by an augmented matrix with two columns, as in the following example:

$$\left[\begin{array}{r|r} 1 & 2 \\ 2 & 5 \\ -1 & -2 \\ 5 & 8 \end{array}\right] \tag{6.25}$$

Gaussian Elimination on this system leads to a contradiction, i.e., the system is inconsistent. Inconsistency means that the system has no exact solution; there is no value for the coefficient w that satisfies the linear equation $wx_i = y_i$ for all the pairs $\langle x_i, y_i \rangle$ specified by (6.25).

Instead of solving $wx_i = y_i$, however, we can solve $wx_i = y'_i$ by computing an estimate of w that yields y'_i values that are as close as possible to the y_i values specified by the system of equations. In other words, we compute w such that it minimizes the error in the inconsistent system of equations. Least Square Estimation is a method for achieving this. It defines the error as the sum of the squares of the differences between the y_i values in the matrix and the estimated y'_i values:

$$e = \sum_i (y_i - y'_i)^2 \tag{6.26}$$

This error function can be simplified by substituting in the linear equation $y'_i = wx_i$:

$$e = \sum_i (y_i - wx_i)^2 = \sum_i (y_i^2 - 2wx_i y_i + w^2 x_i^2) \tag{6.27}$$

To determine the minimum of the error function, we differentiate (6.27) with respect to w:

$$\frac{\partial e}{\partial w} = \sum_i (-2x_i y_i + 2wx_i^2) = -2\sum_i x_i y_i + 2w\sum_i x_i^2 \tag{6.28}$$

The derivative of a function is zero at all points at which the function has a minimum. Hence we can obtain the minimum of the error function by setting the derivative in (6.28) equal to zero:

$$-2\sum_i x_i y_i + 2w\sum_i x_i^2 = 0 \tag{6.29}$$

By resolving this equation to w, we obtain a formula for computing the value of w that minimizes the error:

$$w = \frac{\sum_i x_i y_i}{\sum_i x_i^2} \tag{6.30}$$

To give an example for Least Square estimation, we compute w for the matrix in (6.25):

$$w = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{1\cdot 2 + 2\cdot 5 + (-1)(-2) + 5\cdot 8}{1^2 + 2^2 + (-1)^2 + 5^2} = 1.74 \tag{6.31}$$

Table 6.6: Estimation error for the example matrix in (6.25)

x	y	y'	$(y-y')^2$
1	2	1.74	.068
2	5	3.84	1.346
−1	−2	−1.74	.068
5	8	8.70	.490

We can now test how good the y'_i values obtained by the estimated coefficient $w = 1.74$ fit the y_i values specified by the matrix. We calculate the squared error for each of the items in the matrix, which yields the values in Table 6.6. The mean of this error, the *mean squared error*, is commonly used as a metric for the fit of the model. In the case of our model, we achieve a mean squared error of .493, i.e., each predicted value y' differs on average $\sqrt{.493} = .702$ from the actual value y. This figure indicates how well the least square estimate of the coefficient w fits the system of linear equations.

The method of Least Square Estimation can be generalized to systems of equations with arbitrarily many coefficients. Such systems consist of equations of the form (6.32), generalizing equations with one coefficient of the form $wx_i = y_i$.

$$w_1 x_{i,1} + w_2 x_{i,2} + \ldots + w_j x_{i,j} = y_i \tag{6.32}$$

Equation (6.32) can be written as a sum, yielding (6.33). Note that (6.33) represents a ranking argument for j constraints as defined in (6.12). The coefficient w_j corresponds to the constraint weight $w(C_j)$, the matrix value $x_{i,j}$ corresponds the violation score $v(S_1,C_j) - v(S_2,C_j)$, and the constant y_i represents the acceptability difference ΔH.

$$\sum_j w_j x_{i,j} = y_i \tag{6.33}$$

Therefore, we can use Least Square Estimation on (6.33) to determine the constraint weights $w_1, \ldots, w_j$ that follow from a given set of ranking arguments.

Note that LSE is the method that underlies multiple regression, a standard analytical procedure that allows us to fit a linear equation to a set of data points. There is a close correspondence between our proposal for estimating constraint weights and the estimation of the coefficients of a regression equation:

$$a + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_j x_{i,j} = y_i \tag{6.34}$$

Note that the regression equation contains an additional coefficient a. This constant is not present in our weight equation; we only deal with acceptability *differences*, not with absolute acceptability values. Hence our weight estimation problem reduces the estimation problem to multiple regression with $a = 0$. We will not derive the relevant mathematical background here; the reader is referred to standard textbooks on linear regression (e.g., Edwards 1984; Rietveld and van Hout 1993).

It is important to point out the LOT approach advocated in this chapter differs in crucial respects from linear regression. In contrast to linear regression, LOT is grounded in linguistic theory. Furthermore, LOT provides a more restrictive model of the data than linear regression, as will be argued in Section 7.7.

6.3.6. Data Complexity and Time Complexity

At least n equations are necessary to solve a system of linear equations with n variables. It follows that at least n ranking arguments are needed to determine the constraint weights in an LOT model with n constraints. This means that LSE algorithm is of *linear* data complexity, i.e., its data complexity function is in $O(n)$, where n is the number of constraints to be ranked.

Note that this is an estimate of the best case data complexity of the LSE algorithm. It only holds if all ranking arguments are *informative*, i.e., each ranking argument contributes an equation that is neither redundant (already present in the system) or contradictory (incompatible with another equation). If this is not the case, then the resulting system is overdetermined or underdetermined, which means that additional examples are necessary to find a solution (this solution may only be approximate in the case of an overdetermined system).

A crucial property of an algorithm is its time complexity, i.e., the function that describes how many computation steps the algorithm requires to process an input of a given size. The Gaussian Elimination algorithm is known to be of polynomial time complexity; its time complexity function is in $O(n^3)$. This means that the number of computation steps required to determine the weights for n constraints grows with the cube of n.

We can estimate the time complexity of the Least Square Estimation algorithm for weight estimation proposed in Section 6.3.5.2 as follows. To solve a system of equations with n variables, n derivatives have to be computed. For each derivative, m computation steps are required to compute the sum in (6.30), where m is the number of equations in the system. This entails that the complexity function of LSE is in $O(n \cdot m)$, i.e., LSE is of polynomial time complexity. The time complexity reduces to $O(n^2)$ if we assume $m = n$, i.e., if n informative ranking arguments are available to compute the weights of n constraints (best case data complexity).

A formal proof of the complexity statements for LSE will be left for future research.

6.3.7. Evaluation of Model Fit and Predictivity

Once the parameters of an LOT model have been established using Gaussian Elimination or Least Square Estimation, we need to evaluate how well the model accounts for a given set of ranking arguments. For Least Square Estimation, a standard metric of *model fit* is available in the form of the *mean squared error*, i.e., the mean of the difference between the model's predicted acceptability difference for a given ranking argument, and the actual acceptability difference found experimentally (see the example in Section 6.3.5.2). The mean squared error

of a model, e_μ, can be defined as follows:

(6.35) **Mean Squared Error of a Model**
Let ΔH_i be the acceptability difference for the item i in the data set, and let $\Delta H_i'$ be the acceptability difference predicted by the model for the item i. Then the mean squared error of the model is as given in (6.36), where n is the number of items in the data set.

(6.36) $$e_\mu = \frac{1}{n}\sum_{i=1}^{n}(\Delta H_i - \Delta H_i')^2$$

Note that e_μ is simply the mean of the squared error e, i.e., the quantity that the LSE procedure is designed to minimize (see Equation (6.26)).

We also have to make sure that the model does not *overfit* the data, i.e., that it is not only able to account for the data that was used for parameter estimation, but can generalize to new data of the same type. In linguistic terms, this means that the model is *predictive*. The generalization ability of a model can be tested by applying it to unseen data, i.e., to data that it has not been used to estimate the model parameters. Again, the mean squared error can be used to quantify the model fit on the test data. If the model fit on the training data and on the test data are similar, then it can be concluded that the model is able to generalize.

Standard techniques from machine learning (Mitchell 1997) and computational linguistics (Manning and Schütze 1999) can be used to carry out detailed studies of the behavior of a model on unseen data. These will be discussed in more detail in Section 7.1.3 in the next chapter.

6.3.8. Standard Optimality Theory as a Special Case

An OT grammar can be formulated as a weighted grammar if the constraint weights are chosen in an exponential fashion, so that strict domination of constraints is assured. This observation is due to Prince and Smolensky (1993: 200).

This observation applies to Linear Optimality Theory, as it constitutes a weighted grammar model. We can therefore prove the following theorem:

(6.37) **Subset Theorem**
A Standard Optimality Theory grammar G with the constraint set $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$ and the ranking $C_n \gg C_{n-1} \gg \ldots \gg C_1$ can be expressed as a Linear Optimality Theory grammar G' with the signature $\langle \mathbf{C}, w \rangle$ and the weight function $w(C_i) = b^i$, where $b-1$ is an upper bound for multiple constraint violations in G.

The Subset Theorem can be proved by considering an arbitrary pair of structures S_L and S_W from the same candidate set, where S_L is a loser and S_W is a winner according to the Standard OT grammar G. We now have to show that S_W wins over S_L also in the corresponding

LOT grammar G'. Therefore, we have to show $H(S_W) > H(S_L)$, i.e., that the harmony of the winner is greater than the harmony of the loser (see Definition (6.7)).

We consider the worst case, i.e., the case where $H(S_W)$ is minimal and $H(S_L)$ is maximal. The worst case follows from strict domination in Standard OT: assume that S_L incurs only a single violation of the constraint C_m. Then S_W can incur multiple violations of all constraints C_k with $C_m \gg C_k$ and will still win over S_L.

Based on this observation, we can now compare the harmonies of S_W and S_L. The harmony of S_L is as follows:

$$H(S_L) = -w(C_m) = -b^m \tag{6.38}$$

In other words, the harmony of the loser, $H(S_L)$, is simply the negation of the weight of C_m (as this is the only constraint it violates) which according to (6.37) is equal to b^m. The harmony of S_W according to Definition (6.6) is:

$$H(S_W) = -\sum_{i=1}^{m-1} w(C_i)v(S_W, C_i) \tag{6.39}$$

Note that i ranges over all constraints that are ranked lower than C_m. We substitute the upper bound $b-1$ for $v(S_W, C_i)$, and b^i for $w(C_i)$ and obtain:

$$H(S_W) = -\sum_{i=1}^{m-1} (b-1)b^i \tag{6.40}$$

This equation can be simplified as follows:

$$\begin{aligned} H(S_W) &= -\sum_{i=1}^{m-1} (bb^i - b^i) \\ &= -\sum_{i=1}^{m-1} b^{i+1} + \sum_{i=1}^{m-1} b^i \\ &= -\sum_{i=2}^{m} b^i + \sum_{i=1}^{m-1} b^i \\ &= -(b^m + \sum_{i=2}^{m-1} b^i) + (b + \sum_{i=2}^{m-1} b^i) \\ &= -b^m + b \end{aligned} \tag{6.41}$$

We know that $b > 1$, hence it follows that $H(S_W) > H(S_L)$, which completes the proof of the Subset Theorem in (6.37).

Note that the Subset Theorem holds only if there is an upper bound $b-1$ that limits the number of multiple constraint violations that the grammar G allows. Such an upper bound exists if we assume that the number of violations incurred by each structure generated by G is finite. This assumption seems to be generally true for OT grammars.

6.3.9. Simulation in Optimality Theory with Stratified Hierarchies

The following theorem is complementary to the Subset Theorem that relates LOT and Standard OT:

(6.42) **Superset Theorem**
A Linear Optimality Theory grammar G with the constraint set $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$ and the weight function $w(C_i)$ can be expressed as an Optimality Theory grammar G' with stratified hierarchies.

The proof for the Superset Theorem presupposes *stratified constraint hierarchies* (as introduced by Tesar 1998: 428). A stratum is a subset of the set of constraints. Constraints in the same stratum are not ranked with respect to each other; the strata themselves, on the other hand, are ranked in the usual optimality-theoretic fashion. For the purposes of constraint evaluation, a stratum counts as a composite constraint: the violations of all constraints in a stratum have the same rank. As an example, assume that the constraints C_1 and C_2 are in the same stratum. Then there is no difference between a violation of C_1 and a violation C_2 in terms of constraint evaluation. Also a combined violation of C_1 and C_2 counts the same as a double violation of C_1 or a double violation of C_2. It is important to note that stratified hierarchies are not part of the inventory of Standard OT as defined by Prince and Smolensky (1993). Also, there are a number of alternative proposal for the interpretation of constraint ties (e.g., Müller 1999; Pesetsky 1998).

Assuming stratified hierarchies, the Superset Theorem can be proved as follows. Let all weights defined by $w(C_i)$ in G be positive integers. Then we can define a stratified OT grammar G' with the constraint set $C' = \{C'_1, C'_2, \ldots, C'_n\}$ with the following property: v violations of C_i in G correspond to $w(C_i) \cdot v$ violations of C'_i in G'. Furthermore, assume that all constraints in C' are in the same stratum.

Now the Superset Theorem can be proved by considering an arbitrary pair of structures S_L and S_W from the same candidate set, where S_L is a loser and S_W is a winner according to the LOT grammar G. We now have to show that S_W wins over S_L also in the corresponding stratified OT grammar G'. In other words, we have to show that $S_W > S_L$ in G implies $S_W > S_L$ in G'. By applying the definition of harmony in (6.6), we can transform $S_W > S_L$ in G into (6.43):

$$-\sum_i w(C_i)v(S_W, C_i) > -\sum_i w(C_i)v(S_L, C_i) \quad (6.43)$$

This inequality can be simplified to:

$$\sum_i w(C_i)v(S_W, C_i) < \sum_i w(C_i)v(S_L, C_i) \quad (6.44)$$

By definition, all constraints in G' are in the same stratum, i.e., they are ranked equally. Therefore, the harmony of S_W and S_L in G' is simply determined by the number of constraint violations that S_W and S_L incur. It follows from the definition of G' that S_W incurs a total of

$\sum_i w(C_i)v(S_W, C_i)$ violations in G', while S_W incurs a total of $\sum_i w(C_i)v(S_W, C_i)$ violations in G'. From the inequality in (6.44) it follows that S_W incurs less violations than S_L in G'. This entails that $S_W > S_L$ in G', which completes the proof.

Note that this proof assumes that the weights in G are restricted to positive integers. The restriction to *integers* is not crucial, as a grammar with weights that are rational numbers can be normalized to arrive at integer weights. The assumption that the weights are *positive* means that only constraint violations (not constraint satisfactions) determine the harmony of a structure. This assumption was part of our definition of harmony in Section 6.3.2 (see also the discussion in Section 6.4.4).

6.4. Assessment of Linear Optimality Theory

This section evaluates the LOT model of gradience against the set of criteria in Sections 6.1.1 and 6.1.2. Furthermore, we discuss the relation between LOT and Standard OT, and address some potential criticism related to the fact we use weighted constraints instead of ranked constraints.

6.4.1. Properties of Gradient Linguistic Structures

This section argues that Linear Optimality Theory provides an adequate account for the set of properties of gradient structures that were discusses in Section 6.1.1.

Ranking and cumulativity, arguably the most central properties of gradient structures, are at the heart of the LOT model. Constraint ranking is modeled by the fact that LOT annotates constraints with numeric weights representing the contribution of a constraint to the unacceptability of a structure. Cumulativity is modeled by assuming that the harmony of a structure is computed as the negation of the weighted sum of the weights of the constraint the structure violates.

Once ranking and cumulativity are assumed as part of the LOT model, all other properties of gradient linguistic judgments follow without further stipulations. The ganging up effect is an obvious case; as constraint violations are cumulative, they can gang up. As an example assume that the weight of the constraint C_1 is twice that of the constraint C_2. Then a structure that violates C_1 once will be as ungrammatical as one that violates C_2 twice, i.e., C_2 gangs up against C_1.

Furthermore, LOT allows us to model the distinction between soft and hard constraints. The two constraint types differ in the degree of unacceptability they cause, which in LOT can be captured in terms of constraint weights. Hard constraints are predicted to be associated with high constraint weights, while soft constraints are associated with low weights. This prediction can be tested by applying the Least Square Estimation algorithm to a set of experimentally collected judgments that contain both hard and soft constraint violations. The constraint weights

that LSE determines should show two clusters: soft constraints at the low end of the scale, and hard constraints on the hight end. This property of LOT will be demonstrated in the detail in the next chapter (see Section 7.2).

Another difference between soft and hard constraints is that soft, but not hard constraints can be subject to context effects. In an LOT setting, this means that the weight of a soft constraint varies with context, while the weight of a hard constraint should be stable across all contexts. Again, we can test this prediction by applying LSE to a data set that contains judgments for hard and soft constraint violations, for a number of contexts. If we run LSE separately for each context, then we should find that it the weight estimates for hard constraints are approximately the same across contexts, while the estimates for soft constraints vary. This property of LOT will be demonstrated by a modeling study on gapping Section 7.3 in the next chapter.

Finally, the LOT model is able to accommodate crosslinguistic variation in the same way as Standard OT (see Section 2.6). The assumption is that the same set of constraints applies in all languages, and that crosslinguistic differences are modeled via constraint re-ranking. In a LOT setting, this means the weight of a given constraint will differ from language to language. This prediction can be tested by using LSE to estimate two data sets that contain judgments for the same constraints in two different languages. A study of word order variation will illustrate this property of LOT in Chapter 7.

Moreover, we hypothesized (based on the data presented Chapter 4) that crosslinguistic variation does not affect the type of a constraint. This means that soft constraints are soft across languages, while hard constraints are crosslinguistically hard. In an LOT model, this means that while the weight of a constraint will vary from language to language, the constraint type will stay the same. In other words, a soft constraint should receive a low weight and trigger context effects across languages, while a hard constraint is crosslinguistically associated with a high constraint weight and immune to contextual variation. Chapter 7 will discuss how this hypothesis is implemented in LOT.

6.4.2. Criteria for Models of Gradient Grammaticality

This section evaluates the LOT model of gradience against the criteria proposed in Section 6.1.2.

6.4.2.1. Causal Adequacy

A model of gradience is causally adequate if it provides an explanation as why grammaticality is a gradient, rather than a binary notion. In Linear Optimality Theory, gradience has a double source. Firstly, gradience stems from the fact that constraints are ranked, i.e., some constraint violations trigger a higher degree of ungrammaticality than others. Secondly, gradience is due

to the fact that constraint violations are cumulative, i.e., the ungrammaticality of a structure increases with the number of violations that it incurs.

In Section 6.3.8 we proved the Subset Theorem, which states that Standard OT is a special case of LOT. This means that LOT inherits important features from Standard OT, including the fact that crosslinguistic variation can be accounted for by constraint re-ranking. This means that constraint ranking is independently motivated in LOT, it does not need to be stipulated just to deal with gradient data.

6.4.2.2. Conceptual Adequacy

LOT is designed to provide a maximally accurate model of speakers' intuitions about degrees of well-formedness. As the work on magnitude estimation of linguistic acceptability has shown (Bard et al. 1996; Cowart 1997), speakers are capable of providing reliable *interval* judgments of linguistic acceptability, i.e., they are able to judge how much more or less acceptable a given structure is in relation to another one. The measurements provided by magnitude estimation studies go beyond the categorical or ordinal judgments standardly used in linguistic theory (see also the discussion of judgment elicitation in Chapter 2).

The LOT model reflects this fact by providing a *quantitative* notion of grammaticality. This allows us not only to state that a given structure is more grammatical than another one, but also how much more grammatical it is, i.e., LOT supports statements such as "the grammaticality difference between structure S_1 and structure S_2 is 2.7".

Note that LOT does not allow absolute statements of grammaticality of the form "structure S_1 has a grammaticality of 7.4". Such statements are not supported by magnitude estimation data, which does not provide a fixed scale of grammaticality with definite endpoints (a ratio scale). Also, there are numerous factors that influence the absolute value of grammaticality judgments, including type of instructions, type of fillers used, the modality of the stimuli (spoken or written) (see Section 2.3 for a survey). Therefore the absolute value of the grammaticality judgment for a given sentence is expected to vary from experiment to experiment, while the relative acceptability of two stimuli can be expected to be constant across experiments (this was evidenced, for instance, by the high correlation we achieved in the replication studies in Chapter 5, Experiments 13–15).

6.4.2.3. Empirical Adequacy

The LOT approach achieves a high degree of empirical adequacy. It is fully supported by a suite of experimental studies, providing data for a large range of syntactic phenomena, and for a number of different languages. This is a clear advantage over earlier approaches to gradience (for instance Hayes's (2000) or Müller's (1999)), which only rely on intuitive judgments. Recall that we argued extensively in Section 2.4 that intuitive judgments are inadequate for measuring

gradient acceptability.

Secondly, empirical adequacy requires that a model is able to account for the experimental properties of gradient judgments that we summarized in Section 6.1.1. Most existing models of gradience fall short of this requirement, and OT-based models in particular are unable to model the cumulativity of constraint violations and the ganging up effect (see Section 6.2.3). Linear Optimality Theory, on the other hand, is able to account for all experimental properties of gradient judgments, based on a set of minimal assumptions about the ranking and interaction of constraints. This was discussed in detail in Section 6.4.1.

6.4.2.4. Computational Adequacy

LOT also meets the criterion of computational adequacy. It specifies a *scoring algorithm*, i.e., a way of computing the degree of grammaticality for a given structure. In LOT, the grammaticality of a two structures in defined in terms of their relative harmony. The harmony of a structure is computed as the negation of the weighted sum of the constraint violations that the structure attracts (see Definitions (6.5) and (6.7) for details).

Secondly, the LOT model provides a *training algorithm*, i.e., a way of estimating the constraint weights from a given data set. As we showed in Section 6.3.5, the problem of determining LOT constraint weights reduces to solving a system of linear equations, a familiar mathematical problem. Efficient and well-understood algorithms such as Gaussian Elimination and Least Square Estimation can therefore be applied as training algorithms for LOT.

Note that most of the earlier models of gradience reviewed in Section 6.2 do not include a training algorithm; they all rely on intuitive, manual ways of estimating the model parameters. The only exceptions are the Variable Rule model proposed by Labov (1969) and Cedergren and Sankoff (1974) and Probabilistic OT proposed by Boersma and Hayes (2001). Recall, however, that a number of potential problems with Probabilistic OT were raised in Section 6.2.3.5.

Another computational issue concerns the complexity of LOT. The LOT model offers an attractive data complexity compared with existing approaches: n ranking arguments are sufficient to estimate the weights of n constraints, i.e., the data complexity function of LOT is in $O(n)$ (see Section 6.3.6). Tesar and Smolensky's (1998) learning algorithm has a data complexity $O(n^2)$, while no complexity estimate is available for Boersma's (1998) learning algorithm, for instance.

Also the time complexity of the LOT is attractive; the Gaussian Elimination algorithm is of complexity $O(n^3)$. The LSE algorithm for weight estimation is also polynomial, probably of complexity $O(n \cdot m)$, where n is the number of constraints for which the weights are to be estimated, and m is the number of ranking arguments to be considered.

6.4.2.5. Predictive Adequacy

We defined predictive adequacy as the existence of a systematic way of evaluating the results of the scoring algorithm and the parameter estimation algorithm. This problem can be broken down into the problem of determining the model fit on the training data, and estimating how well the model generalizes, i.e., how well it performs on unseen data.

Both tests can be applied straightforwardly in an LOT framework. In Section 6.3.7 we argued that a standard metric such as the mean squared error can be used to assess how well a given model fits a given data set. Based on this metric for model fit, we can then apply standard machine learning techniques such as crossvalidation to determine the performance of a model on unseen data, and thus assess how well the model is able to generalize. We will demonstrate how this mode of evaluation works for a series of models in Chapter 7.

Note that none of the existing models of gradience has been tested on unseen data. This is a serious shortcoming, as the absence of testing on unseen data leaves open the possibility that a model *overfits* the data, i.e., that it achieves a good fit to the training set, but is unable to generalize to unseen data.

6.4.2.6. Cognitive Adequacy

Throughout this thesis, we have avoided psycholinguistic claims, i.e., claims about the mechanisms involved in human language processing (parsing and generation). The modeling aim of the present thesis is a representational one—the LOT approach is supposed to account for the knowledge that underlies speakers' judgments of the relative well-formedness of an utterance. No direct claims about the *processing* of linguistic knowledge can be derived from LOT. Such claims are typically supported by real-time data (such as eye tracking data or other reaction time measurements), which are not available in the present thesis (but see Bard et al. 1999 and Pechmann et al. 1994 for comparisons of gradient judgments with real-time data).

However, our model has some interesting implications for language *acquisition*. Gradient grammaticality seems to be relevant for several aspects of language development, including first language acquisition, second language acquisition, and language attrition. These phenomena have been shown to involve optionality, i.e., the grammar admits more than one structure as a realization of a given input. Typically, such optional structures are not equally acceptable, but differ in their degree of acceptability, and the relative acceptability of optional forms changes during the course of language development. A formalism like Linear Optimality Theory could be suitable to account for such developmental changes in terms of changes in the constraint weights (developmental re-ranking). We will return to this issue in Section 8.2.3.

We have to bear in mind, however, that the LSE algorithm proposed in this chapter is not meant as an account of how speakers acquire linguistic knowledge. LSE crucially relies on information about the relative acceptability of utterances. Under standard assumptions about

language acquisition, no such data is available to the language learner, as it constitutes a form of negative evidence.

6.4.3. Relationship to Standard Optimality Theory

This section discusses the relationship between the LOT model outlined in Section 6.3 and Standard Optimality Theory (see Section 2.6 for an overview of Standard OT and Section 6.2.3 for a discussion of previous OT-based models of gradience).

Linear Optimality Theory preserves key concepts of Standard Optimality Theory. This includes the fact that constraints are violable, even in an optimal structure. As in Standard OT, LOT avails itself of a notion of constraint ranking to resolve constraint conflicts; LOT's notion of ranking is quantified, i.e., richer than the one in Standard OT. The second core OT concept inherited by LOT is constraint competition. The optimality of a candidate cannot be determined in isolation, but only relative to other candidates it competes with. Furthermore, LOT uses ranking arguments in a similar way as Standard OT. Such ranking arguments work in a competitive fashion, i.e., based on the comparison of the relative grammaticality of two structure in the same candidate set. As in Standard OT, a comparison of structures across candidate sets is not well-defined; two structures only compete against each other if they share the same input.

The crucial difference between LOT and Standard OT is the fact that in LOT, constraint ranks are implemented as numeric weights and a straightforward linear constraint combination scheme is assumed. Standard Optimality Theory can then be regarded as a special case of LOT, where the constraint weights are chosen in an exponential fashion so as to achieve strict domination (see the Subset Theorem in (6.37)).

The extension of Standard OT to LOT allows us to account for the cumulativity of constraint violations, which is something that none of the OT-based models of gradience is able to achieve (see Section 6.2.3). Furthermore, the linear constraint combination schema greatly simplifies the task of determining a constraint hierarchy from a given data set. This problem simply reduces to solving a system of linear equations, a well-understood mathematical problem for which a set of standard algorithms exists, two of which we surveyed here: Gaussian Elimination and Least Square Estimation (see Section 6.3.5). While Gaussian Elimination only returns a result if there is a set of weights that precisely corresponds to the requirements of a set of ranking arguments, Least Square Estimation returns an approximate solution that minimizes the error, i.e., the difference between the acceptability differences predicted by a set of weights and the acceptability differences specified by a set of ranking arguments. The LSE algorithm is therefore suitable for determining constraint weights based on experimental data, which typically contain noise, and will not fit a given model perfectly.

Note that the determination of constraint ranks from data is not a trivial task in Standard OT, and has been the subject of much research (e.g., Tesar and Smolensky 1998; Tesar

1998). The Least Square Estimation constitutes a surprisingly simple solution to the OT learning task. It has the added advantage of being robust to noise in the input, a feature it shares with Boersma's (1998) Gradual Learning Algorithm, which is designed to overcome limitations of Tesar and Smolensky's original approach to OT learning. An intriguing question in this context is if the learning algorithms that we proposed for LOT (Gaussian Elimination and LSE) can also be applied to Standard OT, given that Standard OT is a special case of LOT (as shown in Section 6.3.8). We will return to this question in Section 8.2.

Another advantage is that LOT naturally accounts for optionality, i.e., for cases where more than one candidate is optimal. Under the linearity hypothesis, this simply means that the two candidates have the same harmony score. Such a situation can arise if the two candidates have the same violation profile, or if they have different violation profiles, but the sum of the violation is the same in both cases. No special mechanism for dealing with constraint ties are required in Linear OT. This is an advantage over Standard OT, where the modeling of optionality is less straightforward (see Asudeh 2001 for a discussion). Note also that Tesar and Smolensky's OT learning algorithm cannot cope with optionality in the training data, whereas this poses no problem for the LOT training schemes.)

6.4.4. Relationship to Harmonic Grammar

Recall our discussion of Harmonic Grammar (HG), a precursor of OT formulated as a weighted grammar model implemented in a hybrid connectionist-symbolic architecture (see Section 6.2.1.2).[8] Linear Optimality Theory is similar to HG in that it assumes constraints that are annotated with numeric weights, and that the harmony of a structure is computed as the linear combination of the weights of the constraints it violates.

There are, however, two differences between LOT and HG: (a) LOT only models constraint violations, while HG models both violations and satisfactions; and (b) LOT uses standard least square estimation to determine constraint weights, while HG requires more powerful training algorithms such as backpropagation. We will discuss each of these differences in turn.

LOT requires that all constraints weights have the same sign (only positive weights are allowed, see Section 6.3.2). This amounts to the claim that only constraint violations (but not constraint satisfactions) play a role in determining the grammaticality of a structure. In HG, in contrast, arbitrary constraint weights are possible, i.e., constraint satisfactions (as well as violations) can influence the harmony of a structure. This means that HG allows to define a grammar that contains a constraint C with the weight w and a constraint C' that is the negation of C and has the weight $-w$. In such a grammar, both the violations and the satisfactions of C influence the harmony of a structure.

This point can be illustrated using the constraint PROALIGN, which requires that pronouns precede full NPs (see Experiments 6 and 10). If a structure S violates of PROALIGN,

[8] The present section owes a lot to discussions with Paul Smolensky.

i.e., if it contains a full NP that precedes a pronoun, then this will decrease its harmony of S by a given amount w. In LOT, a satisfaction of PROALIGN (if the structure contains a pronoun that precedes a full NP), however, will not improve its harmony. In Harmonic Grammar, however, we can define an additional constraint PROALIGN$'$ that responds to a satisfaction of PROALIGN, i.e., it decreases the harmony of S by $-w$ if S contains a pronouns that precedes a full NP. Such a constraint PROALIGN$'$ cannot be defined in LOT.

This difference between HG and LOT can also be illustrated with respect the Superset Theorem (see (6.42)). In Section 6.3.9 we proved that an arbitrary LOT grammar can be simulated by an OT grammar with stratified hierarchies. This proof crucially relies on the assumption that all constraint weights are of the same sign. Stratified hierarchies allow us to simulate the addition of constraint violations (they correspond to multiple violations in Standard OT), but they do not allow us to simulate the subtraction of constraint violations (which would be required by constraints that increase harmony). This means that the proof in Section 6.3.9 does not work for grammars that have both positive and negative constraints weights, as they are possible in Harmonic Grammar.[9]

The second difference between HG and LOT concerns parameter estimation. An HG model can be implemented as a connectionist network, and the parameters of the model (the constraint weights) can be estimated using standard connectionist training algorithms. An example is provided by the HG model of unaccusativity/unergativity in French presented by Legendre et al. (1990a,b) and Smolensky et al. (1992). This model is implemented as a multilayer perceptron and trained using the backpropagation algorithm (see Rumelhart, Hinton, and Williams 1986; for a general introduction to connectionist modeling see Bishop 1995).

It is well known that many connectionist models have an equivalent in conventional statistical techniques for function approximation. Multilayer perceptrons, for instance, correspond to a family of non-linear statistical models, as shown by Sarle (1994). (Which non-linear model a given perceptron corresponds to depends on its architecture, in particular the number and size of the hidden layers.) The parameters of a multilayer perceptron are typically estimated using backpropagation or similar training algorithms.

On the other hand, a single-layer perceptron (i.e., a perceptron without hidden layers) corresponds to multiple linear regression, a standard statistical technique for approximating a linear function of multiple variables. The parameters (of both a single-layer perceptron and a linear repression model) can be computed using least square estimation (Bishop 1995). This technique can also be used for parameter estimation for LOT models (see Section 6.3.5.2). Note that LOT can be conceived of as a variant of multiple linear regression. The difference between LOT and conventional multiple linear regression is that parameter estimation is not carried

[9] There seems to be a clear difference between LOT and HG in terms of what type of constraint weights are possible. However, this does not necessarily imply that the two frameworks differ in their generative capacity. As shown by Smolensky et al. (1992), the generative capacity of HG is at least context free. No corresponding proof is available for LOT, hence it remains an open question if HG and LOT differ in generative capacity.

directly on data to be accounted for (the acceptability judgments); rather, a preprocessing step is carried out on the judgment data to compute a set of ranking arguments, which then form the input for the regression. The difference between conventional regression and LOT is explored in more detail in Section 7.7.

To summarize, the crucial difference between HG and LOT is that HG is a non-linear function approximator, while LOT is a linear function approximator, i.e., a variant of linear regression. This means that a different set of parameter estimation algorithms is appropriate for HG and LOT, respectively.

6.5. Conclusions

In this chapter, we established a set of conceptual, empirical, and computational criteria that a model of gradience has to meet. Based on these criteria, we discussed previous models of gradience proposed in the literature, and showed that none of them is fully adequate for modeling gradient grammaticality, in particular in the light of the experimental data we presented in Chapters 3 and 4. This includes pre-OT models such as weighted rule models or probabilistic grammars, and OT-based approaches such as the re-ranking model or Probabilistic OT.

We proposed an alternative model, Linear Optimality Theory, that borrows central concepts such as constraint ranking and constraint competition from Standard OT. Crucially, however, LOT assumes a numeric form of constraint ranking and incorporates a linear constraint combination scheme. This entails that the harmony of a structure is proportional to the weighted sum of the constraint violations it incurs. LOT allows us to define a relative notion of grammaticality, which we argued is adequate for accounting for the properties of gradient judgments identified in Chapters 3 and 4. Furthermore, we proved that Standard OT is a special case of Linear Optimality Theory where constraint weights are set in an exponential fashion to assure strict domination. Linear Optimality Theory, on the other hand, can be simulated in Standard OT if stratified constraint hierarchies are allowed.

Standard OT makes use of ranking arguments to establish constraint hierarchies. We showed that LOT supports ranking arguments similar to the ones used in Standard OT, but allows us to draw on the relative acceptability of suboptimal structures. We demonstrated that a set of ranking arguments for a given candidate set can be reduced to a system of linear equations. Standard algorithms exist for solving such systems of equations, and we showed that two of them, Gaussian Elimination and Least Square Estimation, can be used to estimate the weights for a given set of constraints from a set of ranking arguments. While Gaussian Elimination will only return a result if there is a set of weights that precisely corresponds to the requirements of a set of ranking arguments, Least Square Estimation returns an approximate solution that minimizes the error, i.e., the difference between the acceptability scores predicted by a set of weights and the ones specified by a set of ranking argument. This algorithm is

therefore suitable for determining constraint weights based on experimental data, which typically contains noise. Both algorithms offer attractive complexity properties, viz., linear data complexity and polynomial time complexity. Furthermore, we argued that standard evaluation schemes from machine learning (such as crossvalidation) can be used to evaluate an LOT model, and to assess its capability to generalize to unseen data.

In the next chapter, we demonstrate the validity of Linear Optimality Theory by using it to model a series of experimental results from Chapters 3 and 4. We use Least Square Estimation to derive constraint weights and employ crossvalidation to determine the model fit on the training and test data.

Chapter 7

Applications of the Model

The aim of this chapter is to demonstrate the validity of Linear Optimality Theory as proposed in Chapter 6. We present two types of modeling studies: three small scale proof of concept studies that illustrate how specific properties of gradient data are accounted for in LOT, and a larger, more realistic study that illustrates the interaction of a number of properties of gradient data. The aim of these modeling studies is to show that certain properties of gradient data (the hard/soft distinction, context effects, and crosslinguistic effects) do not have to be stipulated, but follow from the core assumptions of Linear Optimality Theory.

The argumentation in this chapter proceeds as follows. We will first present three modeling studies that demonstrate how the model is able to account for a set of properties of gradient data: in Modeling Study 1, we show how the distinction between hard and soft constraints is modeled, based on extraction data from Experiment 4. Modeling Study 2 then demonstrates how context effects are dealt with by LOT using gapping data from Experiment 8. Modeling Study 3 deals with the modeling of crosslinguistic variation. In Modeling Studies 4 and 5 we present a larger case study that draws together the results from the proof of concept studies and deals with all aspects of gradient data: constraint ranking, hard and soft constraints, context effects, and crosslinguistic variation. This case study provides a detailed model of the word order data for Greek and German from Experiments 6 and 10–12.

Throughout this chapter, Least Square Estimation is employed to determine model parameters (i.e., constraint ranks) from experimentally collected judgment data. Crossvalidation is used to demonstrate that the predictions of a model generalize to unseen data.

At the end of this chapter, we contrast the LOT approach with more conventional analytic tools, viz., analysis of variance and multiple regression. We argue that an LOT approach is to be preferred, as it is grounded in linguistic theory, and provides a more restrictive model of the data.

7.1. Introduction

7.1.1. Obtaining Data for LOT Models

A model of grammatical competition has to provide a way of specifying which candidate structures are involved in the competition. In Optimality Theory, this is achieved by specifying the *input*, i.e., a representation from which a set of competing candidate structures is generated by the generation function *Gen* (see Sections 2.6 and 6.3.1 for more information on the architecture of OT). A number of diverse proposals have been put forward in the OT literature regarding which representations are adequate as inputs. Proposals include predicate argument structures (Legendre et al. 1995), sets of lexemes (Grimshaw 1997), LFG-style f-structures (Bresnan 2000), or syntactic derivations (Müller 1999).

We will keep our assumptions regarding the input as minimal and as theory-neutral as possible, so as to be able to make claims of maximal generality. This in line with the strategy we adopted for postulating constraint sets in Chapters 3 and 4, where we opted for a descriptive, surface-oriented formulation of our constraints.

We assume that the input specifies the set of lexical items that are to be realized by the candidates. A set of candidates is then generated from the input by the generation function *Gen* such that all structures incorporate the lexical items specified by the input, possibly augmented with functional elements such as determiners, pronouns, and clitics. Accent placement (relevant for Experiment 12 only) is also added by *Gen*.

The input also specifies the Information Structure of the utterance, i.e., each constituent in a candidate structure is marked as either focus or ground. The consequence is that a given candidate set contains only candidates with the same Information Structure (uttered in the same context). This assumption will be relevant for the modeling of context effects in Modeling Studies 2, 4, and 5.

7.1.2. Training LOT Models

In Section 6.3.5 we showed that the problem of determining the ranking of a set of constraints from a set of acceptability judgments reduces to the problem of solving a system of linear equations. We introduced Least Square Estimation (LSE) as a method for achieving this task in a way that minimizes the mismatch between the judgments predicted by the model and the ones found experimentally.

Throughout the present chapter, we will we use LSE to determine the constraint weights for a given model, and show that LSE is able to achieve a high fit between the model and the experimental data, and also that it yields intuitive constraint rankings for the data it is applied on.

Here are the details of how LSE will be applied in the modeling studies reported in this chapter:

1. First, we compute the means of the experimentally obtained judgments for each structure that was tested in the experiment to be modeled. These means constitute the data to be accounted for by the model.

2. Then we identify the candidate sets, i.e., sets of structures that share the same input. The criteria for this step depend on the type of linguistic phenomenon we are dealing with; one key criterion throughout this chapter is that the structures with the same context have to be in the same candidate set. (Recall that a candidate set is a set of structures that compete against each other; structures that share the same candidate set are typically displayed in the same tableau in the OT literature.)

3. Then all ranking arguments are computed for all candidate sets that were identified. Recall from Section 6.3.4 that a ranking argument is the comparison of two structures, based on the difference in violation profile and the difference in acceptability. Note that ranking arguments can only be applied to structures in the same candidate set.

4. The set of all ranking arguments for all candidate sets for a given experiment constitutes the data set to be modeled. We run the LSE algorithm is run to compute the constraint weights for the set of constraints assumed in this model.

5. The model fit is estimated by computing the mean squared error (see Section 6.3.7). The data is split into a training set and a test set to perform crossvalidation (see Section 7.1.3) in order to make sure that the model does not overfit the data.

6. Separate runs of the LSE algorithm on subsets of the data can be carried out to determine context effects on the constraint weights (see Modeling Study 2 for details).

These steps will be illustrated in more detail by Modeling Studies 1–5.

7.1.3. Testing LOT Models

To assess the quality of a model, we have to quantify how well the model fits the data set it was trained on. A suitable metric for the fit of an LOT model is e_μ, the mean squared error (MSE), i.e., the mean of the difference between the model's predicted harmony difference for a given ranking argument, and the actual acceptability difference found experimentally (see Section 6.3.7 for details). The MSE is particularly useful for assessing the performance of the Least Square Estimation algorithm, as this algorithm is designed to minimize the squared error. This means that the MSE directly reflects the success of training using Least Square Estimation.

Throughout this chapter, we will also employ another metric for the model fit, viz., the accuracy of a model, which is defined as follows:

(7.1) **Accuracy of a Model**

Let ΔH_i be the acceptability difference for the item i in the data set, and let $\Delta H_i'$ be the

acceptability difference predicted by the model for the item i. Then the accuracy of the model is as given in (7.2), where n is the number of items in the data set and $hit(x)$ is a function that returns 1 if $x \leq \delta$, and 0 otherwise.

$$A = \frac{1}{n} \sum_{i=1}^{n} hit(|\Delta H_i - \Delta H_i'|) \tag{7.2}$$

The accuracy measures how often the model correctly predicts an acceptability difference in the data set: we count a hit if the predicted acceptability difference does not diverge from the actual acceptability difference by more than a given threshold δ. For the purpose of the present chapter, we will use as δ the mean standard deviation of the items in the data set.[1] In other words, we count a hit if the predicted score does not diverge by more than one standard deviation from the actual score.

The accuracy metric makes it possible to compare the model fit of different models, even if the models have been trained on different data sets. Throughout this chapter, we will give both the MSE (which is specific to a given data set), and the accuracy. All accuracies will be expressed as percentages.

Using the MSE and the accuracy metric, we can not only assess the model fit on the training data, but we can also test for overfitting. A model overfits if it provides a good fit on the data set it was trained on, but only a poor fit on a set of related, but unseen data. Overfitting means that the model is unable to generalize to new instances of the same problem and thus fails to capture the regularities in the data.

Tests for overfitting are available in the form of standard crossvalidation techniques in machine learning (Mitchell 1997) or computational linguistics (Manning and Schütze 1999). The following techniques are the most common ones:

- **Held-Out Data** This approach involves randomly splitting the data set into two sets, the training set that is used to estimate the parameters of the model, and the test set that is used to test the model. Then the model fit is computed on both the test set and the training set; a good model fit on the test set indicates that the model is able to generalize to unseen data, i.e., does not overfit the training data.

  The disadvantage of the held-out data approach is that a fairly large data set has to be used; the test set should be about 10% of the overall data set; if the data set is too small, no meaningful results can be achieved when testing the model.

- ***k*-fold Crossvalidation** This approach is a generalization of the held-out data approach. The data set is randomly partitioned in k subsets. The model is tested on one of these subsets, after having been trained on the remaining $k-1$ subsets. This is procedure is repeated k times such that each of the subset serves once as test set and

[1] The standard deviations for all experimental data are given in Appendix C.

$k-1$ times as part of the training set. Based on the training and testing results, average values for the model fit can be computed.

The *k*-fold crossvalidation approach has the advantage of being applicable also to fairly small data sets, as in effect the whole data set is used for testing. Also, we obtain average values for the model fit on the training and the test data, i.e., confidence intervals can be computed. Typically, a value of $k = 10$ is used in the literature.

- **Leave One Out** This method is an instance of *k*-fold crossvalidation where *k* is set to the size of the data set. This means that we train on all items of the training set, leaving out only one item, on which the model is then tested. This procedure is then repeated *k* times and the average model fit is computed.

  The advantage of leave one out is that it is even more suitable for small data sets than standard *k*-fold crossvalidation. An obvious disadvantage is that a large number of training and test runs have to be carried out

In the following, we will use *k*-fold crossvalidation for all our models, and set $k = 10$. We report the mean squared error and the accuracy on both the training set and the test set. Our MSE and accuracy figures are averaged over 10 training runs, and are given with 95% confidence intervals.

7.2. Modeling Study 1: Soft and Hard Constraints

In this section, we will apply the LOT model to the judgment data for extraction from picture NPs from Experiment 4 (see Section 3.5). This proof of concept study will illustrate how the LOT model works and how it is able to account for the distinction between soft and hard constraints.

7.2.1. Constraints and Candidate Sets

The constraint set for this modeling study is the same one that was used to discuss the results of Experiment 4. It includes three soft constraints on picture NP extraction, viz., DEFINITENESS (DEF), VERBCLASS (VERB), and REFERENTIALITY (REF). We also tested three hard constraints on extraction: INVERSION (INV), RESUMPTIVE (RES), and AGREEMENT (AGR). Section 3.5.1 describes these constraints in more detail and provides relevant examples.

As far as the candidate set is concerned, we assume an input representation that contains a set of lexical categories (such as nouns, verbs, adjectives, etc.) to be realized by the structures in a given candidate set. The generation function *Gen* then augments this input with functional categories (such as determiners and complementizers), and adds features such as [±DEF], [±EX], and [±REF]. We assume that the feature [±DEF] marks the definiteness of a

Table 7.1: Violation profile for extraction data (Experiment 4)

wh_i have you V NP of t_i	RES	AGR	INV	VERB	REF	DEF
wh_i[+REF] have you V[−EX] NP[−DEF] of t_i						
wh_i[+REF] have you V[−EX] NP[+DEF] of t_i						*
wh_i[−REF] have you V[−EX] NP[−DEF] of t_i					*	
wh_i[−REF] have you V[−EX] NP[+DEF] of t_i					*	*
wh_i[+REF] have you V[+EX] NP[−DEF] of t_i				*		
wh_i[+REF] have you V[+EX] NP[+DEF] of t_i				*		*
wh_i[−REF] have you V[+EX] NP[−DEF] of t_i				*	*	
wh_i[−REF] have you V[+EX] NP[+DEF] of t_i				*	*	*
wh_i[+REF] has you V[−EX] NP[−DEF] of t_i		*				
wh_i[+REF] have you V[−EX] NP[−DEF] of him$_i$	*					
wh_i[+REF] has you V[−EX] NP[−DEF] of him$_i$	*	*				
wh_i[+REF] you have V[−EX] NP[−DEF] of t_i			*			
wh_i[+REF] you has V[−EX] NP[−DEF] of t_i		*	*			
wh_i[+REF] you have V[−EX] NP[−DEF] of him$_i$	*		*			
wh_i[+REF] you has V[−EX] NP[−DEF] of him$_i$	*	*	*			

noun phrase, [±EX] indicates whether a verb is existential or not, and [±REF] marks a *wh*-phrase as referential. (See Keller 1997 for a more detailed discussion of the input adequate for an OT analysis of extraction from picture NPs.)

Recall that in Experiment 4, we tested 16 different sentence types, incurring either between zero and three soft violations or between zero and three hard violations. All these sentences types can be generated from the same input, they only differ in terms of their feature specification and in their lexical realization (and in terms of the surface order of the constituents in the case of INVERSION violations). Therefore, all 16 structures compete with each other in the same candidate set, which is given in Table 7.1, together with the violation profiles of the candidates (note that this tableau contains only 15 structures as the null violation condition was included twice in the experiment).

7.2.2. Ranking Arguments and Constraint Ranks

The ranking for a given constraint set can be determined empirically based on a set of relevant acceptability judgments. To achieve this, we compute all the ranking arguments generated by the acceptability judgments; the resulting set of ranking arguments can then be used to derive the constraint ranks. In LOT, a ranking argument is based on the comparison of the violation profiles of two structures from the same candidate set. The acceptability difference between these two structures serves as evidence for the ranking of the constraints violated by the two structures (see Section 6.3.4 for details).

As outlined in Section 7.1.2, we will compute all ranking arguments for all candidate

Table 7.2: Constraint weights for extraction data (Experiment 4)

fold	RES	AGR	INV	VERB	REF	DEF	e_μ(train)	e_μ(test)	A(train)	A(test)
1	.2530	.1896	.1417	.0236	.0167	.0024	.0095	.0100	96.29	100.00
2	.2495	.1917	.1443	.0209	.0048	.0153	.0095	.0099	96.29	100.00
3	.2448	.2024	.1503	.0265	.0194	.0081	.0088	.0176	97.22	83.33
4	.2511	.1879	.1398	.0221	.0063	.0053	.0101	.0048	96.29	100.00
5	.2478	.1813	.1544	.0320	.0103	−.0012	.0090	.0150	97.22	91.66
6	.2607	.1879	.1444	.0272	.0020	.0066	.0091	.0135	98.14	91.66
7	.2501	.1797	.1422	.0216	.0127	−.0051	.0099	.0067	97.22	100.00
8	.2458	.1789	.1508	.0253	.0054	.0022	.0098	.0072	96.29	100.00
9	.2530	.1779	.1471	.0269	.0071	.0006	.0097	.0085	98.14	91.66
10	.2484	.1750	.1436	.0171	−.0029	−.0097	.0093	.0131	97.22	100.00
mean	.2505	.1853	.1459	.0244	.0082	.0025	.0095	.0107	97.04	95.83
95% CI	.0103	.0190	.0107	.0096	.0155	.0162	.0009	.0093	1.67	13.48

sets that are contained in a given set of acceptability judgments. Recall that for a candidate set of k elements, there are $(k^2-k)/2$ ranking arguments (see Section 6.3.4). The judgment data from Experiment 4 provide only one candidate set, the one depicted in Table 7.1. By computing all ranking arguments generated by this candidate set, a set of 120 data points is obtained. (Modeling Studies 2, 4, and 5 will rely on training data consisting of more than one candidate set.)

Now we can use Least Square Estimation to compute a set of constraint weights from this set of ranking arguments. Recall that LSE is an algorithm that allows us to estimate the weights in a way that minimizes the mismatch between the acceptability differences predicted by the model and the ones found experimentally (see Section 6.3.5.2 for details).

The aim of the present study is to determine how well an LOT model with the constraints DEF, VERB, REF, INV, RES, and AGR fits the experimental data. However, we also want to test whether the model is predictive, i.e., whether it is able to generalize to unseen data. As explained in Section 7.1.3, this can be achieved by carrying out crossvalidation. In the present study, we applied ten-fold crossvalidation, i.e., the data (the set of ranking arguments) was randomly divided into ten test sets of equal size, and the constraint weights were determined from the remainder of the data. Table 7.2 reports the constraint weights for each of the ten folds, and Figure 7.1 graphs the average constraint weights with 95% confidence intervals.

The highest constraint weight of .2505 is obtained for RES, the constraint against resumptive pronouns. The constraints AGR and INV also achieve relatively high weights of .1853 and .1459, respectively. These high weights reflect the fact that RES, AGR, and INV are hard constraints, i.e., a violation of these constraints causes serious unacceptability. As the confidence interval of these three weights fail to overlap, we conclude that the weights are distinct, and we can derive the overall constraint ranking of RES $\gg$ AGR $\gg$ INV corresponds to the ranking that was obtained in Experiment 4 on the basis of an ANOVA carried out directly on the judgment data. This finding demonstrates that the LSE algorithm can determine plausible con-

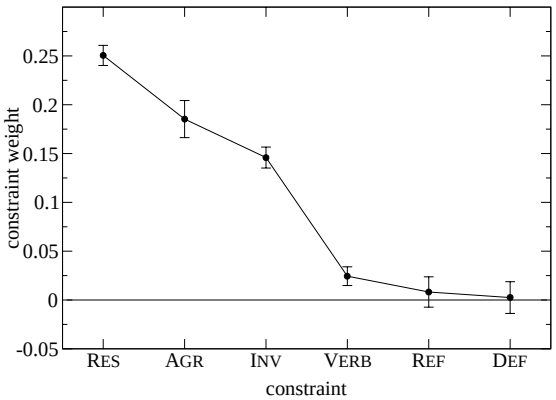

Figure 7.1: Constraint weights for extraction data (Experiment 4)

straint weights based on ranking arguments. In contrast to ANOVA, LOT achieves this by using a technique that is motivated by linguistic theory as it is based on concepts from Optimality Theory.

The constraint weights of VERB, REF, and DEF were estimated at .0244, .0082, and .0025, respectively. That these constraints are soft constraints is illustrated by the fact that their weights are considerably lower than the ones of the hard constraints RES, AGR, and INV. Also note that the weights of VERB, REF, and DEF are fairly similar to each other (the confidence intervals overlap), i.e., they receive approximately the same constraint rank. To summarize, the modeling results suggest that the LSE method yields plausible constraint ranks that reflect the soft/hard distinction.

7.2.3. Model Fit and Predictions

Table 7.2 reports the model fit on each of the folds, and the average model fit computed over all ten folds. The average mean squared error on the training data, $e_\mu(\text{train})$, is .0095. The average accuracy on the training data, $A(\text{train})$, is 97.04%, which means that 97.04% of the time prediction of the model is within one standard deviation of the actual acceptability difference in the training set (the mean standard deviation for the experimental data was .2075). This indicates that the LOT model achieves a good fit on the experimental data.

The average mean squared error on the test data, $e_\mu(\text{test})$, is .0107, the average accuracy on the test data, $A(\text{test})$, is 95.83%. This means that the model is able to generalize to unseen instances of extraction data, as the model fit on the test data is close to that on the training data.

7.2.4. Conclusions

This proof of concept study showed that the LOT model can be applied to experimentally collected extraction data. A high model fit was obtained and it was shown that the model generalizes to unseen data. The LOT model generated a set of constraint ranks that are plausible (in the sense of being compatible with other analyses on the same data).

Crucially, this study illustrated how LOT can capture the distinction between hard and soft constraints by assigning high constraint weights to hard constraints and low weights to soft constraints. This illustrates how the soft/hard dichotomy emerges from the core assumptions of LOT and does not need to be stipulated separately.

7.3. Modeling Study 2: Context Effects

This section applies LOT to the contextualized judgment data for gapping constructions that were obtained as part of Experiment 8. As in the previous section, the aim of this modeling study is to provide a proof of concept for LOT; this time we demonstrate how context effects can be captured in the LOT framework.

7.3.1. Constraints and Candidate Sets

The constraint set in this study is the same as in Experiment 8. We use the Minimal Distance Principle (MINDIS), the Tendency for Subject-Predicate Interpretation (SUBJPRED) and the Requirement for Simplex-Sentential Relationship (SIMS), all proposed by Kuno (1976) (see Section 4.2.1 for details).

We assume that the input representation for gapping is the full sentence structure, including both conjuncts without gaps. The generation function *Gen* then deletes subconstituents of this input, leaving behind remnants. These gapped structures then compete against each other in a candidate set. We assume that the generation function will also add a feature [±CONTR] that indicates if a verb is a subject control verb. This feature is important for evaluating the constraint SUBJPRED: if the remnants include an object NP and a VP, then only a subject control verb in the matrix will allow the NP and the VP to be interpreted as subject and predicate. (See Section 4.2.1 for details.)

Experiment 8 elicited judgments for eight different types of gapped structures. In these structures, gapping occurred in the right conjunct, leaving behind exactly two remnants. All eight structures that were part of this experiment can be generated from the same input, i.e., they compete in the same candidate set. Table 7.3 lists the candidates and gives their violation profiles.

Recall that the structures in Experiment 8 were presented in four difference contexts: null context, neutral context, felicitous context, and non-felicitous context. In the neutral con-

Table 7.3: Violation profile for gapping data (Experiment 8)

NP V NP [V NP] and NP V NP [V NP]	SIMS	MINDIS	SUBJPRED
NP V[+CONTR] NP [V NP] and _ _ NP [V NP]			
NP V[+CONTR] NP [V NP] and _ _ NP [_ NP]	*		
NP V[−CONTR] NP [V NP] and NP _ _ [V NP]		*	
NP V[−CONTR] NP [V NP] and NP _ _ [_ NP]	*	*	
NP V[−CONTR] NP [V NP] and _ _ NP [V NP]			*
NP V[−CONTR] NP [V NP] and _ _ NP [_ NP]	*		*
NP V[+CONTR] NP [V NP] and NP _ _ [V NP]		*	*
NP V[+CONTR] NP [V NP] and NP _ _ [_ NP]	*	*	*

text, the gapped sentence was prefixed by an all-focus question like *What happened?*. In the felicitous context, the gap represented contextually given information, while the remnant constituted new information. In the non-felicitous context, this situation was reversed.

As outlined in Section 7.1.1, we assume that the input for the competition also specifies the context of the utterance in which the utterance is to be realized. For the gapping data, this means that we have to assume a separate candidate set for each of the four context; candidate structures are not allowed us to compete across contexts. All four candidate sets contain the eight structures given in Table 7.3.

7.3.2. Ranking Arguments and Constraint Ranks

The gapping data provides four candidate sets (one for each context), as outlined above. For each of these sets, 28 ranking arguments can be computed, i.e., we obtain a data set that contains a total of 112 ranking arguments. We used this data set to determine the constraint weights for the three constraints SUBJPRED, MINDIS, and SIMS. As in Modeling Study 1, the data were split into ten separate test sets to carry out ten-fold crossvalidation. The resulting constraint weights are listed in Table 7.4 and graphed in Figure 7.2.

We found a high constraint weight of .1638 for the constraint SIMS, which was classified as a hard constraint in Experiment 8. The soft constraints MINDIS and SUBJPRED, on the other hand, received rather low constraints weights of .0874 and .0433, respectively. The confidence intervals for the three constraints weights do not overlap, which means that the weights are distinct, and we conclude that the ranking is SIMS $\gg$ MINDIS $\gg$ SUBJPRED. This is the same ranking that was obtained in Experiment 8 on the basis of an ANOVA carried out directly on the judgment data. This provides further evidence for the claim that our the LSE algorithm computes plausible constraint weights based on OT-style ranking arguments. It also confirms that LOT is able to model the soft/hard distinction in an intuitive fashion.

Recall that in Experiment 8 we concluded that the soft constraint MINDIS was subject to context effects, whereas the hard constraint SIMS was immune to contextual variation. The

Table 7.4: Constraint weights for gapping data (Experiment 8)

fold	SIMS	MINDIS	SUBJPRED	e_μ(train)	e_μ(test)	A(train)	A(test)
1	.1664	.0822	.0443	.0037	.0036	100.00	100.00
2	.1616	.0822	.0410	.0035	.0057	100.00	100.00
3	.1621	.0880	.0441	.0038	.0030	100.00	100.00
4	.1665	.0843	.0415	.0037	.0032	100.00	100.00
5	.1626	.0936	.0426	.0036	.0042	100.00	100.00
6	.1642	.0898	.0430	.0036	.0047	100.00	100.00
7	.1616	.0882	.0450	.0037	.0037	100.00	100.00
8	.1643	.0871	.0429	.0038	.0029	100.00	100.00
9	.1643	.0913	.0444	.0034	.0059	100.00	100.00
10	.1640	.0865	.0438	.0039	.0017	100.00	100.00
mean	.1638	.0874	.0433	.0037	.0039	100.00	100.00
95% CI	.0041	.0082	.0029	.0002	.0029	0.00	0.00

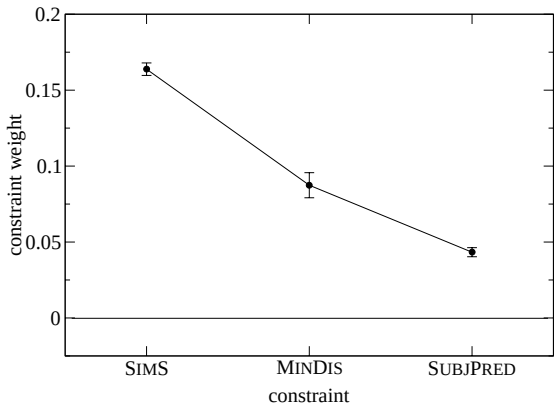

Figure 7.2: Constraint weights for gapping data (Experiment 8)

status of SUBJPRED was less clear, only limited context effects were observed for this constraint. In an LOT setting, context effects should be reflected in constraint weights. If a given constraint is context-dependent, then this means that its weight is higher in some contexts than in others, i.e., we should observe *context-specific re-ranking* for this constraint.

To test this hypothesis, we conducted separate LSE runs for the four contexts that were included in the gapping data. As each of the data sets was small (28 items), no crossvalidation was performed, but we trained on the whole data set. This entails that no testing on unseen data was possible, and no confidence intervals are available for the context-specific weights.

The constraint weights for each context are graphed in Figure 7.3. We observed only a small context-specific difference in the weight of SIMS. This is in line with predictions (and with the findings of Experiment 8), as SIMS is a hard constraint, and thus is expected to be context-independent. For the soft constraint MINDIS, however, context-specific re-ranking can be observed: the weight of MINDIS drops from .1009 in the non-felicitous context to .0233 in the felicitous context (the null context and the neutral context behave like the non-felicitous

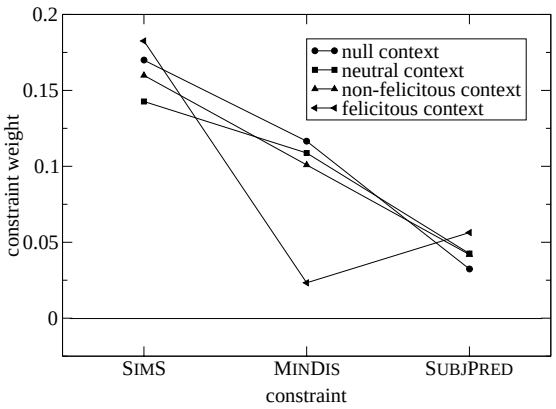

Figure 7.3: Context effects for gapping data (Experiment 8)

context). On the other hand, only weak re-ranking effects were obtained for SUBJPRED. This is compatible with the findings of Experiment 8, where SUBJPRED showed context effects only under very specific conditions (viz., for single constraint violations compared with the null violation condition).

To summarize, the present modeling study provides a clear-cut case of context-specific re-ranking: the ranking MINDIS $\gg$ SUBJPRED is reversed to SUBJPRED $\gg$ MINDIS in a felicitous context.

7.3.3. Model Fit and Predictions

Table 7.4 reports the model fit on each fold, and the average model fit over all ten folds. For the training data, we find an average e_μ of .0037, which indicates a very good fit for the gapping model. This is confirmed by the fact that the accuracy is 100% for all ten folds (the mean standard deviation for the experimental data was .2226).

The average mean squared error on the test data is .0039, which indicates only a slight decrease compared to the model fit on the training data. The accuracy on the test set is again 100%. We conclude that our model is very good at generalizing to unseen instances of gapping data.

7.3.4. Conclusions

We presented another proof of concept study for the LOT model, deriving the constraint weights for a set of gapping judgments using LSE and showing that these weights correspond to intuitively plausible constraint ranks. The resulting model reflected the soft/hard dichotomy: hard constraints were assigned high weights, soft ones received low weights. A high model fit was achieved both on the training data and on unseen data, indicating that the model is predictive.

The main aim of this modeling study was to show how LOT can capture context effects, i.e., the fact that the influence of certain soft constraints is context-dependent. We demonstrated that effects are modeled in LOT as context-specific re-ranking: the weight of a context-dependent constraint varies from context to context. For the gapping data, the LSE algorithm correctly generates context-specific re-ranking for context-dependent, but not for context-independent constraints. Note that the re-ranking approach to context effects follows naturally from the core assumptions of LOT, i.e., from the fact that constraints are weighted and that constraint violations are cumulative. No additional stipulations are necessary to capture context effects in LOT.

7.4. Modeling Study 3: Crosslinguistic Variation

This section provides another proof of concept study for the LOT model. The aim of this study is to illustrate the crosslinguistic aspects of the model. We show how LOT, just like Standard OT, is able to account for crosslinguistic variation in terms of constraint re-ranking.

7.4.1. Constraints and Candidate Sets

The experiments presented in Chapters 3 and 4 provide no suitable data for illustrating crosslinguistic variation directly. The one crosslinguistic study that we conducted (on word order preferences in Experiments 6 and 10–12) only contained one constraint that was tested in two languages (GROUNDALIGN), which is not sufficient as an example for crosslinguistic re-ranking in LOT. In the present section, we will therefore discuss a hypothetical data set that is rich enough to demonstrate how LOT deals with crosslinguistic variation. This data set deals with word order variation and draws on the results obtained in Experiments 6 and 10–12.

We assume a constraint set that contains four constraints on linear order: NOMALIGN, DATALIGN, VERBFINAL, and VERBINITIAL. NOMALIGN states that nominative NPs have to precede non-nominative NPs, while DATALIGN specifies that dative NPs have to precede accusative NPs. The constraint VERBFINAL requires the verb to occur in sentence final position, while VERBINITIAL requires the verb to occur in sentence initial position. (See Section 3.7.1 for a more extensive discussion of these constraints.)

Consider the following data on word order variation in the subordinate clause in German. These data are analogous to the data investigated in Experiment 6, where we dealt with the order of NP complements in the subordinate clause in German. Six permutations of the subject, object, indirect object are possible in the subordinate clause:

Table 7.5: Violation profile and hypothetical ratings for word order data (Greek and German)

{V,S,I,O}	VERBFIN	VERBINI	NOMAGN	DATAGN	German	Greek
SIOV		*			12	7
SOIV		*		*	11	6
ISOV		*	*		10	6
OSIV		*	*	*	9	5
IOSV		*	**		8	5
OISV		*	**	*	7	4
VSIO	*				7	12
VSOI	*			*	6	11
VISO	*		*		5	11
VOSI	*		*	*	4	10
VIOS	*		**		3	10
VOIS	*		**	*	2	9

(7.3) a. **VSIO:** Ich hoffe, dass der Mann dem Vater den Kaffee gibt.
I hope-1SG that the man-ACC the father-DAT the coffee-ACC give-3SG
"I hope the man gives the father the coffee."

b. **VSOI:** Ich hoffe, dass der Mann den Kaffee dem Vater gibt.

c. **VISO:** Ich hoffe, dass dem Vater der Mann den Kaffee gibt.

d. **VOSI:** Ich hoffe, dass den Kaffee der Mann dem Vater gibt.

e. **VIOS:** Ich hoffe, dass dem Vater den Kaffee der Mann gibt.

f. **VOIS:** Ich hoffe, dass den Kaffee dem Vater der Mann gibt.

The order VSIO fails to violate any constraints, while VSOI violates DATALIGN, and VISO violates NOMALIGN. The order VOSI violates both NOMALIGN and DATALIGN once, while VIOS violates NOMALIGN twice, and VOIS violates NOMALIGN twice and DATALIGN once.

In Experiment 6, we found that a violation of NOMALIGN is more serious than a violation of DATALIGN. Violations of VERBFINAL were investigated separately in Experiment 10, where both verb final and verb initial subordinate clauses were tested. The results showed that a VERBFINAL violation is more serious than a NOMALIGN violation.

For the present modeling study, we will assume a candidate set that includes both verb final and verb initial structures for all six permutations, resulting in a set of 12 structure with the violation profiles in Table 7.5. This table also lists an acceptability score for each structure. While these scores were not derived experimentally, they are compatible with the results of Experiment 6 and 10, i.e., they reflect the relative strength of NOMALIGN, DATALIGN, and VERBFINAL violations as observed in these experiments. Note that a violation of VERBINITIAL does not play a role for the word order data for German; VERBINITIAL is violated in verb final clauses without a reduction in acceptability.

For crosslinguistic purposes, it is interesting to compare the data for German in (7.3)

with data for subordinate clauses in Greek. Main and complement clauses generally allow a relatively free word order in Greek, as discussed in Section 4.6.1. An exception is provided, however, by complement clauses in the subjunctive mood governed by matrix verbs such as *eplizo* "hope", *epithimo* "wish", *diatazo* "order", and *apagorevo* "forbid". These subordinate clauses have to be verb initial, and the order of the NP complements of the subordinate verb is more restricted. Consider the examples in (7.4), which are parallel to the data for German in (7.3).

(7.4) a. **VSIO:** Eplizo na dosi o adras tou patera ton kafe.
hope-1SG SUBJ give-3SG the man-ACC the father-GEN the coffee-ACC
"I hope the man gives the father the coffee."

b. **VSOI:** Eplizo na dosi o adras ton kafe tou patera.

c. **VISO:** Eplizo na dosi ton kafe o adras tou patera.

d. **VOSI:** Eplizo na dosi tou patera o adras ton kafe.

e. **VIOS:** Eplizo na dosi ton kafe tou patera o adras.

f. **VOIS:** Eplizo na dosi tou patera ton kafe o adras.

We will assume the same constraint set for Greek that was used for German, containing the constraints NOMALIGN, DATALIGN, VERBFINAL, and VERBINITIAL. Note that this requires a slight modification of the constraint DATALIGN; this constraint now has to be formulated as: dative or genitive NPs have to precede accusative NPs. (This formulation is based on the assumption that the genitive in Greek has the same function as the dative in German.)

There is no experimental data available for subordinate sentences such as the ones in (7.4). However, native speaker intuitions suggest that a violation of either NOMALIGN or DATALIGN leads to a small decrease in acceptability. There seems to be no difference in the amount of unacceptability induced by NOMALIGN and DATALIGN. A violation of VERBINITIAL, however, triggers strong unacceptability. A violation of VERBFINAL, on the other hand, fails to have an effect on acceptability.

We assume the set of 12 candidates structures in Table 7.5, i.e., the same candidate set as for German. This table lists the violation profiles for examples such as the ones in (7.4) and also contains a set of acceptability scores for Greek. Note that these scores were not derived experimentally, but they do reflect the intuitions of our informants. This type of data is sufficient in the present context, were the aim is to provide a proof of concept, rather than a rigorous linguistic analysis.

All the structures in Table 7.5 compete with each other, as they can be generated from the same set of input representations (containing the constituents V, S, I, and O). Note that context was not taken into account in the present modeling study (but see Modeling Studies 4 and 5 for an account of word order preferences that includes context effects).

Table 7.6: Constraint weights for hypothetical word order data

	VERBFIN	VERBINI	NOMAGN	DATAGN
German	5.0000	0.0000	2.0000	1.0000
Greek	0.0000	5.0000	1.0000	1.0000

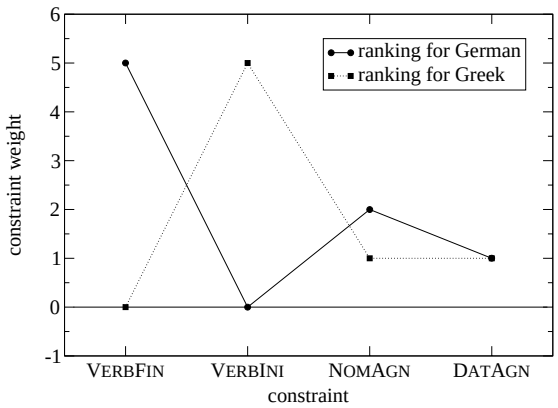

Figure 7.4: Constraint weights for hypothetical word order data

7.4.2. Ranking Arguments and Constraint Ranks

The data in Table 7.5 provide two candidate sets with 12 items each, one for German and one for Greek. Based on this, a set of 66 ranking arguments can be computed for each language, which can then be used to estimate the weights for the constraints NOMALIGN, DATALIGN, VERBFINAL, and VERBINITIAL with the help of the Least Square Estimation algorithm. The resulting constraint weights are listed in Table 7.6 and graphed in Figure 7.4.

For the German data, a high constraint weight of 5.0 was found for the constraint VERBFINAL which reflects the fact that non-final verb order leads to strong unacceptability. On the other hand, the weight of VERBINITIAL was zero, indicating that this constraint fails to play a role in German. Furthermore, the constraint NOMALIGN received a weight of 2.0, whereas the weight of DATALIGN was estimated at 1.0. This amounts to an overall constraint hierarchy of VERBFINAL $\gg$ NOMALIGN $\gg$ DATALIGN $\gg$ VERBINITIAL, which is what we expect based on the results of Experiments 6 and 10.

For the Greek data, we find a high weight of 5.0 for VERBINITIAL, while the weight of VERBFINAL was estimated at zero. Both the constraints NOMALIGN and DATALIGN received a weight of 1.0. The resulting ranking VERBINITIAL $\gg \{$NOMALIGN, DATALIGN$\} \gg$ VERBFINAL reflects the intuition that a violation of VERBINITIAL has a strong affect on acceptability in Greek, while NOMALIGN and DATALIGN trigger only weak unacceptability, and VERBFINAL fails to have any effect on acceptability.

Two instances of *crosslinguistic re-ranking* of constraints can be studied by comparing the constraint weights for the German and the Greek data. First, consider the constraints

VERBFINAL and VERBINITIAL that regulate verb order. Both of these constraints are hard constraints, i.e., they lead to strong unacceptability when violated. However, only one of the constraints is *active* in each language: VERBFINAL in German and VERBINITIAL in Greek. (See Prince and Smolensky 1993: 107 for a definition of active constraints that is applicable to this situation.)

The training algorithm detects the inactivity of a constraint because a violation of an inactive constraint is not associated with an acceptability difference. This is reflected in the constraint weight by the fact that this inactive constraint receives a weight of zero.[2] The second case of crosslinguistic re-ranking occurs for the constraints NOMALIGN and DATALIGN. In German, we find NOMALIGN $\gg$ DATALIGN, whereas in Greek, both constraints receive the same rank. Note that both constraints are active in both languages, i.e., they have a non-zero constraint weight. Also, both constraints receive low constraint weights in both languages, i.e., they can be classified as soft constraints. However, the relative weights of NOMALIGN and DATALIGN differ crosslinguistically; this is how re-ranking is modeled in LOT.

Note that this example also illustrates the hypothesis about crosslinguistic effects that was put forward in Chapter 4. We hypothesized that crosslinguistic re-ranking cannot change the type (soft or hard) of a constraint, i.e., there are no constraints that are soft in one language and hard in another (see Section 4.1.2 for details). This means that a soft constraint will receive a low constraint weight across languages, while a hard constraint will receive a high constraint weight across languages. This hypothesis is instantiated by the present example: the constraints VERBFINAL and VERBINITIAL are crosslinguistically hard, while the constraint NOMALIGN and DATALIGN are crosslinguistically soft. On the other hand, these constraints are subject to crosslinguistic re-ranking, even though their constraint type is stable across languages.

Note that in evaluating this hypothesis, we have to discount cases where a given constraint is inactive in a language; the zero weight of VERBINITIAL in German, for instance, does not imply that this constraint is soft in this language, it merely indicates that it is inactive in German.

Recall Modeling Study 2, where contextual re-ranking was discussed as a second criterion for the soft/hard distinction: soft constraints are subject to context-specific re-ranking, while hard constraints are immune to contextual re-ranking effects. This leads to the general observation that the re-ranking behavior of a constraint can serve as a diagnostic of its constraint type.

[2] A constraint will also receive a weight of zero if there is no evidence for it in the training set, i.e., if all ranking arguments exhibit a violation difference of zero for this constraint. Therefore, if the LSE returns a weight of zero for a given constraint, we have to inspect the training data to determine whether this zero is due to lack of evidence or constitutes a genuine zero, i.e., reflects the fact that a violation of this constraint fails to trigger any acceptability effects.

7.4.3. Model Fit and Predictions

Both data sets for this model were constructed such that a perfect model fit can be achieved, hence we find $e_\mu = 0$, i.e., there is no divergence between the predicted and the observed acceptability differences. Note that no crossvalidation was carried out on the hypothetical data, hence only one run of the estimation algorithm is reported in Table 7.6.

7.4.4. Conclusions

This proof of concept study illustrated how crosslinguistic variation is captured in LOT. The study showed that the crosslinguistic re-ranking of two constraints is modeled in LOT as a change in the relative weight of the constraints. We also discussed a case where a constraint is inactive in a given language, i.e., the constraint has no effect on acceptability and is assigned a constraint weight of zero by the LSE algorithm. Note that these facts about crosslinguistic variation are a consequence of core assumptions of LOT and do not have to be stipulated separately (just like the facts about contextual variation presented above).

Furthermore, this study provided a more precise version of the hypothesis on crosslinguistic variation that was introduced in Chapter 4: crosslinguistic re-ranking can change the rank, but not the type (soft or hard) of a constraint. This hypothesis was instantiated in the model presented in this section. Note that the model was based on hypothetical, but plausible data on word order variation in German and Greek.

Drawing together results on context-specific re-ranking and crosslinguistic re-ranking, we arrived at the general hypothesis that the re-ranking behavior of a constraint can be used as a diagnostic for its constraint type (i.e., to determine whether the constraint is soft or hard). This predicts that other instances of constraint re-ranking should exist that also correlate with constraint type. This hypothesis will be explored further in Section 8.2.

7.5. Modeling Study 4: Word Order in German

Modeling Studies 1–3 were limited proof of concept studies designed to illustrate how LOT captures certain aspects of gradient data, viz., hard and soft constraints, context effects, and crosslinguistic variation. Modeling Studies 4 and 5 go beyond this and show how a detailed LOT model of a given linguistic phenomenon (word order) can be obtained, based on data from a series of magnitude estimation experiments. This case study also illustrates how an LOT approach can yield results that are consistent across experiments. We chose word order for this case study because of the richness of the data (spanning four experiments) and because of its crosslinguistic dimension.

The modeling study reported in this section uses LSE to determine weights for the constraints on German word order investigated in Experiments 6 and 10. This automatically

derived hierarchy can then be compared to the constraint rankings that we hypothesized in Section 4.5.5 based on the experimental findings.

7.5.1. Constraints and Candidate Sets

The constraint set for the German data builds on the one used in Modeling Study 3 for our LOT account of crosslinguistic variation. It combines constraints that were used in the previous modeling study, viz., NOMALIGN (nominative NPs precede non-nominative NPs), DATALIGN (dative NPs precedes accusative NPs), and VERBFINAL (verbs are sentence final), with the additional constraints PROALIGN (full NPs precede pronominalized NPs) and GROUNDALIGN (ground NPs are peripheral). (See Section 3.7.1 for a more extensive discussion of these constraints.)

We first model the data from Experiment 6. This experiment investigated how the constraints NOMALIGN, DATALIGN, and PROALIGN determine the word order in verb final clauses containing a nominative subject, an accusative object, and a dative object. Following the assumptions about the input described in Section 7.1.1, we work on the hypothesis that all permutations of subject, direct object, and indirect object compete with each other, based on the assumption that they are generated from the same input, and therefore end up in the same candidate set. The input representation is an unordered set of constituents from which the generation function *Gen* computes all possible permutations. As outlined in Section 7.1.1, *Gen* adds feature specifications when it generates a candidate set; examples include the definiteness feature [±DEFINITE] and the feature [±CONTR] for subject control verbs (see Modeling Studies 1 and 2). In the case of word order data, *Gen* adds the feature [±PRO] that indicates whether an NP is pronominalized no not.

Such a conception of the input has the consequence that structures with pronominalized and non-pronominalized NPs compete in the same candidate set. This is a desirable consequence, as it allows us to model the role that pronominalization plays in discourse. In a given discourse context, certain realizations of the input will be dispreferred because they contain (or fail to contain) a pronoun. For instance, we expect a sentence to incur a penalty if it realizes ground information (i.e., information that is contextually given) as a full NP instead of pronominalizing it. A focused NP, on the other hand, cannot normally be pronominalized, and has to be realized as a full NP. Such effects can only be modeled if both structures (the one with the full NP and the one with the pronoun) compete in the same candidate set. In such a setting, we assume that the information structure (the context) of an utterance is part of the input specification based on which the generation function computes candidate structures that may contain full NPs, pronouns, or clitics.

Experiment 6 used non-contextualized stimuli. However, as argued in Chapter 4, even a null context contains implicit information structural assumptions; we expect it to behave like an all focus context. We are therefore justified in assuming that all structures included in Ex-

Table 7.7: Violation profile for German word order data (Experiment 6)

{V,S,I,O}	NOMAGN	PROAGN	DATAGN
SIOV			
SOIV			*
ISOV	*		
IOSV	**		
OSIV	*		*
OISV	**		*
S_{pro}IOV			
S_{pro}OIV			*
IS_{pro}OV	*	*	
IOS_{pro}V	**	**	
OS_{pro}IV	*	*	*
OIS_{pro}V	**	**	*
SI_{pro}OV		*	
SOI_{pro}V		**	*
I_{pro}SOV	*		
I_{pro}OSV	**		
OSI_{pro}V	*	**	*
OI_{pro}SV	**	*	*
SIO_{pro}V		**	
SO_{pro}IV		*	*
ISO_{pro}V	*	**	
IO_{pro}SV	**	*	
O_{pro}SIV	*		*
O_{pro}ISV	**		*

periment 6 compete in the same candidate set. Six word orders were included (all permutations of S, I, and O), and each order was realized either with three full NPs or with two full NPs and a pronoun. This yields a total of 24 word orders, which are listed in Table 7.7. This table also gives the violation profile with respect to the constraints NOMALIGN, DATALIGN, and PROALIGN.

Experiment 10 extended Experiment 6 by investigating contextualized stimuli, with four different contextual specifications: null context, all focus context, S focus context, and O focus context. This allows us to assess the effect of GROUNDALIGN, which constrains the position of ground constituents (and requires them to be peripheral, i.e., sentence initial or final). Furthermore, Experiment 10 manipulated the verb position, both verb final and verb initial sentences were tested, which makes it possible to assess the effect of VERBFINAL, the requirement that subordinate clauses are verb final. Four word orders were tested, viz., SOV, OSV, VSO, and VOS (note that only transitive verbs were included in Experiment 10, while Experiment 6 dealt with ditransitive verbs). The object and the subject could be realized either

Table 7.8: Violation profile for German word order data, all focus context (Experiment 10)

{V,S,O}	VERBFIN	NOMAGN	PROAGN	GAGN
SOV				
OSV		*		
VSO	*			
VOS	*	*		
$S_{pro}OV$				
$OS_{pro}V$		*	*	
$VS_{pro}O$	*			
VOS_{pro}	*	*	*	
$SO_{pro}V$			*	
$O_{pro}SV$		*		
VSO_{pro}	*		*	
$VO_{pro}S$	*	*		
$S_{pro}O_{pro}V$				
$O_{pro}S_{pro}V$		*		
$VS_{pro}O_{pro}$	*			
$VO_{pro}S_{pro}$	*	*		

Table 7.9: Violation profile for German word order data, S focus context (Experiment 10)

{V,S,O}	VERBFIN	NOMAGN	PROAGN	GAGN
SOV				*
OSV		*		
VSO	*			
VOS	*	*		*
$SO_{pro}V$			*	*
$O_{pro}SV$		*		
VSO_{pro}	*		*	
$VO_{pro}S$	*	*		*

as a full NP or pronominalized (in the null context condition, we also included stimuli where both NPs are pronominalized). This yields a total of 16 candidates in the candidate set for the null context condition, and a total of eight candidates in each of the candidate sets for the context conditions. The resulting candidate sets for the null context is given in Table 7.8, and Tables 7.9 and 7.10 contain the candidate sets for the S focus and the O focus context. We omit the candidate set for the null focus context, which is identical to the one for the null context, but omits all structures that contain a pronominalized subject.

Table 7.10: Violation profile for German word order data, O focus context (Experiment 10)

{V, S, O}	VERBFIN	NOMAGN	PROAGN	GAGN
SOV				
OSV		*		*
VSO	*			*
VOS	*	*		
$S_{pro}OV$				
$OS_{pro}V$		*	*	*
$VS_{pro}O$	*			*
VOS_{pro}	*	*	*	

Table 7.11: Constraint weights for German word order data (Experiment 6)

fold	NOMAGN	PROAGN	DATAGN	e_μ(train)	e_μ(test)	A(train)	A(test)
1	.1837	.0972	−.0013	.0160	.0118	97.18	100.00
2	.1857	.0983	−.0041	.0145	.0252	97.98	92.85
3	.1819	.0893	.0059	.0152	.0200	98.39	92.59
4	.1809	.0982	.0039	.0160	.0121	97.17	100.00
5	.1829	.0937	−.0016	.0162	.0105	97.17	100.00
6	.1821	.0965	.0016	.0161	.0106	97.18	100.00
7	.1859	.0964	.0007	.0154	.0174	97.58	96.42
8	.1832	.0965	−.0050	.0149	.0216	97.59	96.29
9	.1842	.0975	−.0064	.0152	.0191	97.58	96.42
10	.1853	.0963	.0027	.0161	.0114	97.17	100.00
mean	.1836	.0960	−.0004	.0156	.0160	97.50	97.46
95% CI	.0038	.0061	.0093	.0013	.0121	.96	6.84

7.5.2. Ranking Arguments and Constraint Ranks

The candidate set for Experiment 6 contains 24 candidates, which yields a total of 276 ranking arguments. We used this data set to compute the weights of the constraints NOMALIGN, DATALIGN, and PROALIGN with Least Square Estimation. As in our previous modeling studies, the data were split into test and training sets using ten-fold crossvalidation. The resulting constraint weights are listed in Table 7.11 and graphed in Figure 7.5.

The results show a high constraint weight of .1836 for NOMALIGN, the constraint that requires nominative NPs to precede non-nominative NPs. The constraint PROALIGN requiring pronouns to precede full NPs receives a lower weight of .0960. This indicates that a NOMALIGN violation triggers stronger unacceptability than a PROALIGN violations. The confidence intervals of the weights of NOMALIGN and PROALIGN do not overlap, hence we conclude that the weights are distinct, i.e., that the ranking NOMALIGN $\gg$ PROALIGN holds. Note that this finding contrasts with the results of our analyses for Experiment 6, where we failed to find a difference between in the ranking of the two constraints. However, this conclusion was based only on a comparison of single constraint violations (following our operational

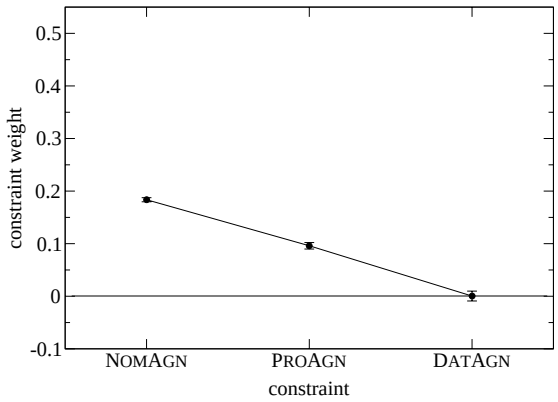

Figure 7.5: Constraint weights for German word order data (Experiment 6)

definition of constraint ranking in Section 3.1.2). The present results takes into account both single and multiple violations of NOMALIGN and PROALIGN, which leads the LOT model to give a higher weight to NOMALIGN. While a simple ANOVA based on single violations was sufficient to correctly establish the ranking for Experiments 4 and 8 (see Modeling Studies 1 and 2), it seems that for more complicated experimental data such as the one in the present model, we have to take into account both single and multiple violations, something which the LOT estimation algorithm is designed to do.

A surprising observation concerns the constraint DATALIGN, which specifies that dative NPs have to precede accusative ones. This constraint does not seem to play a role in our model; it receives an average constraint weight of close to zero (it is even assigned a negative constraint weight in some folds, which means that a violation actually improves acceptability instead of reducing it). An explanation for this phenomenon was already provided when we discussed the results of Experiment 6 in Section 3.7.6. The effect of DATALIGN seems to be limited to non-pronominalized stimuli; once we are dealing with pronouns, the relative position of dative and accusative NPs becomes irrelevant. This observation can be tested in an LOT setting by splitting the candidate set in two subsets, one containing non-pronominalized stimuli and one with pronominalized stimuli. We generated ranking arguments from these two candidate sets and obtained a set of 15 ranking arguments for the candidates without pronouns, and 153 for the ones with pronouns. Then we carried out separate runs of LSE for the two sets, yielding the constraint weights in Figure 7.6 (no crossvalidation was conducted).

The weights for NOMALIGN for both data sets are comparable, i.e., .2120 for the non-pronominalized set, and .1741 for the pronominalized set. The weight of PROALIGN can only be computed for the pronominalized set, where it takes on a value of .1227. Although the numeric weights of NOMALIGN and PROALIGN in the pronominalized set are different from the ones obtained for the full data set, the ranking NOMALIGN $\gg$ PROALIGN remains the same. However, there is a sharp increase in the weight of DATALIGN for non-pronominalized

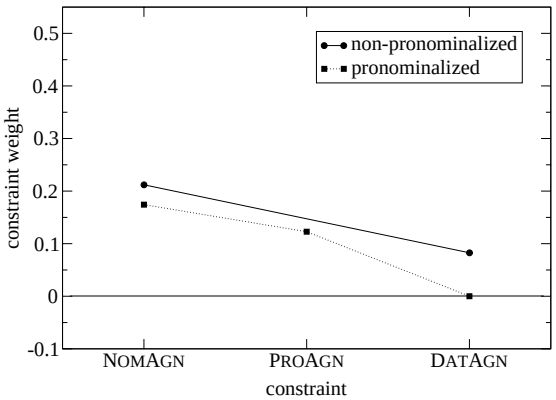

Figure 7.6: Pronominalization effects for German word order data (Experiment 6)

Table 7.12: Constraint weights for German word order data (Experiment 10)

fold	VERBFIN	NOMAGN	PROAGN	GAGN	e_μ(train)	e_μ(test)	A(train)	A(test)
1	.4166	.1468	.1014	.0474	.0194	.0186	95.65	95.00
2	.4203	.1387	.0967	.0456	.0184	.0274	96.19	95.00
3	.4264	.1380	.0902	.0286	.0196	.0165	94.53	100.00
4	.4250	.1419	.0979	.0395	.0191	.0215	96.19	85.00
5	.4249	.1381	.0917	.0419	.0188	.0237	96.17	90.47
6	.4201	.1317	.0807	.0483	.0194	.0192	94.56	95.00
7	.4190	.1434	.0895	.0410	.0193	.0191	95.65	95.00
8	.4146	.1324	.0887	.0356	.0188	.0238	95.08	95.23
9	.4277	.1444	.0954	.0360	.0196	.0169	95.10	100.00
10	.4209	.1439	.0792	.0222	.0198	.0154	95.08	95.23
mean	.4226	.1400	.0912	.0386	.0193	.0202	95.42	94.60
95% CI	.0098	.0116	.0164	.0190	.0009	.0087	1.46	9.91

NPs; DATALIGN receives a weight of .0826 for this data set, compared with a weight of $-.0281$ for the pronominalized data set. This confirms the conclusion we arrived at in Experiment 6: the effect of DATALIGN is limited to sentences with three full NPs, as soon as one of the NPs is realized as a pronoun, the effect disappears.

This finding serves as an example of how LOT can model effects that only concern a part of the data. We split the data into subsets based on theoretically motivated criteria (such as pronominalization) and then compare the constraint weights for these subsets. Note that this is another instance of the re-ranking technique that also served to detect context effects (see Modeling Study 2) and crosslinguistic effects (see Modeling Study 3).

We will now turn to the data from Experiment 10, which provides four candidate sets, three of them with eight candidates (for all focus, S focus, and O focus context), and one with 16 candidates (for the null context). This yields a total of 204 ranking arguments. This data set was used to compute the constraint weights for VERBFINAL, NOMALIGN, PROALIGN, and GROUNDALIGN. The resulting constraint weights are given in Table 7.12 and Figure 7.7.

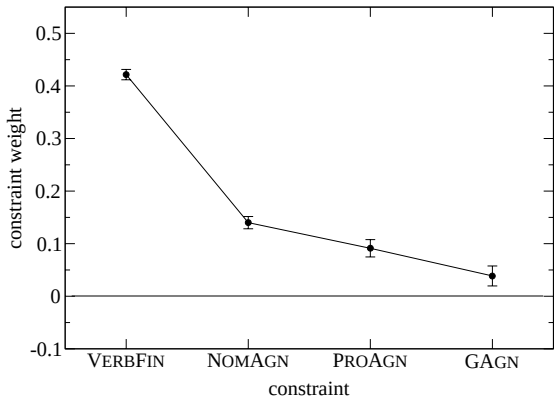

Figure 7.7: Constraint weights for German word order data (Experiment 10)

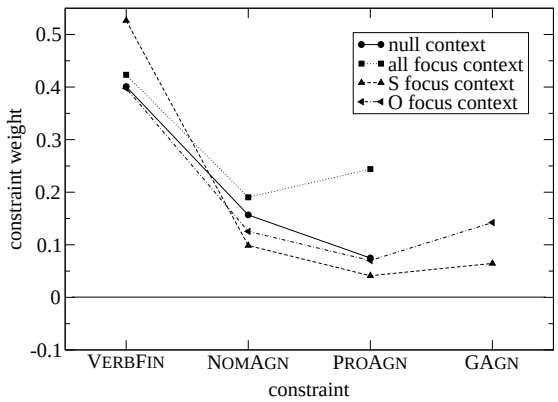

Figure 7.8: Context effects for German word order data (Experiment 10)

The highest ranking of .4226 is obtained for VERBFINAL, a hard constraint. The other constraints are soft constraints and receive comparatively low rankings, viz., .1400 for NOMALIGN, .0912 for PROALIGN and .0386 for GROUNDALIGN. The confidence intervals for all four constraints fail to overlap, hence we conclude that the constraint weights are distinct, and we arrive at the overall constraint hierarchy VERBFINAL $\gg$ NOMALIGN $\gg$ PROALIGN $\gg$ GROUNDALIGN. This hierarchy is compatible with the ranking derived in Section 4.5 based on an ANOVA on the experimental data. Note that the modeling study yields the additional ranking NOMALIGN $\gg$ PROALIGN that was not detected by the ANOVA (in line with the model for Experiment 6). The reason for this is that the ANOVA was only computed on single violations and only took into account the non-contextualized data (due to limitations of the factorial design), while the parameters of the LOT model were estimated on the whole data set.

To demonstrate how LOT can model context effects on word order, we carried out separate LSE runs for the four contexts that were included in the data set. As in the gapping model reported in Modeling Study 2, soft constraints can be subject to context-specific re-

ranking, while hard constraints should be immune to re-ranking. Figure 7.8 confirms this prediction (note that some context fail to provide evidence for certain constraints; these constraints are omitted from the graph). The highly ranked hard constraint VERBFINAL exhibits only small variations in constraint weight from context to context. The soft constraints PROALIGN shows a clear pattern of context-specific re-ranking: in the all focus contexts, PROALIGN outranks NOMALIGN, while in all other context, the ranking PROALIGN $\gg$ NOMALIGN holds. GROUNDALIGN is also subject to context-specific re-ranking: in the S focus context its weight is similar to that of PROALIGN, while in the O focus context the ranking GROUNDALIGN $\gg$ PROALIGN holds. Note that no clear re-ranking effects can be observed for NOMALIGN, even though it is a soft constraint: context-specific re-ranking is not a necessary property of soft constraints.

7.5.3. Model Fit and Predictions

Tables 7.11 and 7.12 report the model fit for the two modeling studies on the German word order data. The model for Experiment 6 shows an excellent performance with an average MSE of .0156, and an average accuracy of 97.50% (the mean standard deviation for the experimental data was .2763). The model also generalizes well, as is evidenced by an MSE of .0160 and an accuracy of 97.46% on the test data.

The model for Experiment 10 performs slightly worse than the first model. Here, the MSE on the training data is .0193, and the accuracy is 95.42% (the mean standard deviation for the experimental data was .2722). Again, we find a good ability to generalize: the MSE on the test data is .0202, which corresponds to a an accuracy of 94.60%,

7.5.4. Conclusions

The LOT model for the German word order data presented in this section illustrated that an LSE-based approach delivers constraint rankings that closely match those obtained using linguistic analysis. It also demonstrated that LOT is able to generate constraint rankings that are consistent across experiments. Furthermore, the LOT model brought to light additional linguistic facts, such as the ranking NOMALIGN $\gg$ PROALIGN that we failed to detect in our discussion of Experiments 6 and 10.

7.6. Modeling Study 5: Word Order in Greek

The study reported in this section extends Modeling Study 4 and accounts for the Greek word order data from Experiments 11 and 12. Recall that no constraint ranking could be determined experimentally for these data, because the set of constraints under investigation did not match the set of factors in the experimental design (see Section 4.6.3.3). However, an LOT approach

Table 7.13: Violation profile for Greek word order data, null context and all focus context (Experiment 11)

{V, S, O}	VAGN	GAGN	DOUAGN
SVO			
OVS			*
VSO			
VOS			
SOV	*		*
OSV	*		*

and Least Square Estimation make it possible to automatically derive a constraint hierarchy for these data based on ranking arguments, as the present modeling study will show.

7.6.1. Constraints and Candidate Sets

The constraint set for the Greek data is the one we used in our discussion of Experiments 11 and 12 and is described in detail in Section 4.6.1. It consists of the phonological constraints ACCENTFOCUS (accented constituents are focussed) and ACCENTALIGN (accent falls on the rightmost constituent), the constraints on clitic doubling DOUBLEALIGN (preverbal objects are doubled) and DOUBLEGROUND (doubled constituents are ground), and the word order constraints GROUNDALIGN (ground constituents are peripheral) and VERBALIGN (the verb must not be right peripheral). Note that the constraint GROUNDALIGN is the same as the one that was used for the German data in Experiment 10.

The assumptions about the input that underlie this model are essentially the same as for the German data. The input representation is an unordered set of constituents from which the generation function *Gen* generates all permutations of subject, object, and verb. As before, we assume that *Gen* also determines the pronominalization of the candidates; for the models in this modeling study this means that *Gen* can add clitic doubling to a candidate. Furthermore, the generation function can also enrich the candidates with accent; this assumption is relevant for our model of the data from Experiment 12, which was based on spoken stimuli. As in previous modeling studies, the input specification is assumed to include the context of an utterance, which has the consequence that only candidates that realize the same information structure compete with each other.

Experiment 11 investigated the constraints VERBALIGN, DOUBLEALIGN, and GROUNDALIGN and included all six permutations of S, O, and V (without clitic doubling). Table 7.13 lists the candidates and their violation profiles for the null context and the all focus context (recall that we assume that a null context behaves like an all focus context). Tables 7.14–7.16 give the candidates and violation profiles for the S focus, O focus, and V focus context, respectively.

Table 7.14: Violation profile for Greek word order data, S focus context (Experiment 11)

{V,S,O}	VAGN	GAGN	DOUAGN
SVO			
OVS			*
VSO			
VOS		*	
SOV	*	*	*
OSV	*		*

Table 7.15: Violation profile for Greek word order data, O focus context (Experiment 11)

{V,S,O}	VAGN	GAGN	DOUAGN
SVO			
OVS			*
VSO		*	
VOS			
SOV	*		*
OSV	*	*	*

Experiment 12 extended Experiment 11 by investigating the additional constraints DOUBLEGROUND, ACCENTFOCUS, and ACCENTALIGN. The experiment comprised three word orders (SVO, OVS, and VSO), either in a clitic doubled and in a non-doubled version, and with either subject or object accent. This yields a total of 12 structures for each of the five contexts (null context, all focus, S focus, O focus, and S focus). Tables 7.17–7.20 list the violation profiles for the resulting candidate sets. (Note that accent is indicated by capitalization, doubling is abbreviated by "cl".)

7.6.2. Ranking Arguments and Constraint Ranks

Each of the five candidate sets for Experiment 11 is made up of six candidates, yielding 15 ranking arguments. We used the resulting overall data set of 60 ranking arguments to estimate the

Table 7.16: Violation profile for Greek word order data, V focus context (Experiment 11)

{V,S,O}	VAGN	GAGN	DOUAGN
SVO			
OVS			*
VSO		*	
VOS			
SOV	*		*
OSV	*		*

Table 7.17: Violation profile for Greek word order data, null context and all focus context (Experiment 12)

{V,S,O}	ACCF	DOUG	GAGN	DOUAGN	ACCAGN
Svo					*
ovS				*	
vSo					*
svO					
Ovs				*	*
vsO					
Sclvo					*
oclvS					
clvSo					*
sclvO		*			
Oclvs		*			*
clvsO		*			

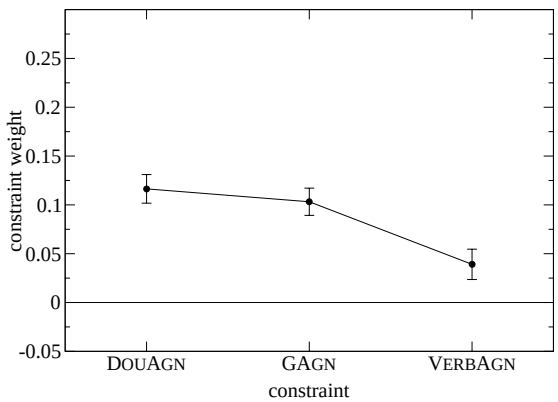

Figure 7.9: Constraint weights for Greek word order data (Experiment 11)

weights for the constraints VERBALIGN, DOUBLEALIGN, and GROUNDALIGN with Least Square Estimation, again using ten-fold crossvalidation to test the performance of the model on unseen data. The resulting constraint weights are listed in Table 7.21 and graphed in Figure 7.9.

The results demonstrate that the constraints DOUBLEALIGN and GROUNDALIGN are roughly of equal importance (the confidence intervals overlap), receiving constraint weights of .1164 and .1032, respectively. The constraint VERBALIGN seems to have only a limited effect on acceptability; it receives a low weight of .0391. This amounts to an overall constraint ranking of {DOUBLEALIGN, GROUNDALIGN} $\gg$ VERBALIGN. All three constraints can be classified as soft constraints, based on the low constraint weights they receive. GROUNDALIGN therefore seems to be a soft constraint in both German and Greek. This is in line with the hypothesis that

Table 7.18: Violation profile for Greek word order data, S focus context (Experiment 12)

{V,S,O}	ACCF	DOUG	GAGN	DOUAGN	ACCAGN
Svo					
ovS				*	
vSo					
svO	*				
Ovs	*			*	
vsO	*				
Sclvo					
oclvS					
clvSo					
sclvO	*				
Oclvs	*				
clvsO	*				

Table 7.19: Violation profile for Greek word order data, O focus context (Experiment 12)

{V,S,O}	ACCF	DOUG	GAGN	DOUAGN	ACCAGN
Svo	*				
ovS	*			*	
vSo	*		*		
svO					
Ovs				*	
vsO			*		
Sclvo	*	*			
oclvS	*	*			
clvSo	*	*	*		
sclvO		*			
Oclvs		*			
clvsO		*	*		

crosslinguistic re-ranking can change the ranking, but not the type of a constraint.

Now consider Figure 7.10 which depicts the result of performing a separate LOT analysis for each of the five contexts tested in Experiment 11 (note that evidence for GROUNDALIGN is only available in narrow focus contexts). All three constraints show a pattern that is characteristic of soft constraints, i.e., they exhibit context-specific re-ranking. This is consistent with the low overall weight that all three constraint receive, which is also a feature of soft constraints. An interesting contextual pattern emerges for DOUBLEALIGN: this constraint receives a high weight in the null context and an intermediate weight in the V focus context, but seems to be of fairly low importance in the other three contexts. This is consistent with the observations about the context-dependence of DOUBLEALIGN that we made in our discussion of Experiment 11 in Section 4.6.6. The LOT approach gives us a means of quantifying such

Table 7.20: Violation profile for Greek word order data, S focus context (Experiment 12)

{V,S,O}	ACCF	DOUG	GAGN	DOUAGN	ACCAGN
Svo	*				
ovS	*			*	
vSo	*		*		
svO	*				
Ovs	*			*	
vsO	*		*		
Sclvo	*				
oclvS	*				
clvSo	*		*		
sclvO	*				
Oclvs	*				
clvsO	*		*		

Table 7.21: Constraint weights for Greek word order data (Experiment 11)

fold	DOUAGN	GAGN	VAGN	e_μ(train)	e_μ(test)	A(train)	A(test)
1	.1281	.0912	.0284	.0057	.0049	100.00	100.00
2	.1113	.1075	.0379	.0056	.0050	100.00	100.00
3	.1247	.1038	.0304	.0053	.0078	100.00	100.00
4	.1174	.1078	.0360	.0059	.0028	100.00	100.00
5	.1166	.1098	.0420	.0055	.0058	100.00	100.00
6	.1148	.1061	.0397	.0058	.0035	100.00	100.00
7	.1081	.0983	.0453	.0056	.0054	100.00	100.00
8	.1193	.1004	.0408	.0057	.0044	100.00	100.00
9	.1135	.0979	.0380	.0053	.0083	100.00	100.00
10	.1096	.1089	.0520	.0047	.0132	100.00	100.00
mean	.1164	.1032	.0391	.0056	.0061	100.00	100.00
95% CI	.0146	.0139	.0155	.0006	.0068	.00	.00

contextual variation.

We will now discuss the modeling results for Experiment 12. The model is based on five candidate sets, each comprising 12 candidates, yielding 66 ranking arguments. The resulting overall data set of 330 ranking arguments was used to estimate the weights for the new constraints ACCENTFOCUS, ACCENTALIGN, DOUBLEGROUND, as well as for the constraints DOUBLEALIGN and GROUNDALIGN that were already included in Experiment 12. As usual, Least Square Estimation with ten-fold crossvalidation was employed for training and testing. The resulting constraint weights are listed in Table 7.22 and graphed in Figure 7.11.

The two constraints ACCENTFOCUS and DOUBLEGROUND receive weights of .1890 and .1785, respectively. As the two weights are very similar (the confidence intervals overlap), we conclude that ACCENTFOCUS and DOUBLEGROUND have the same constraint rank. Note also that the high constraint weights are characteristic of hard constraints (re-

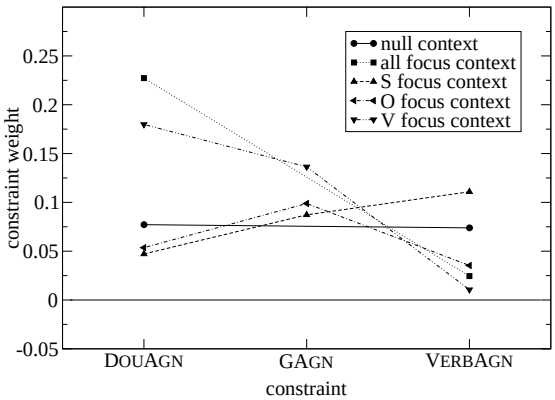

Figure 7.10: Context effects for Greek word order data (Experiment 11)

Table 7.22: Constraint weights for Greek word order data (Experiment 12)

fold	ACCF	DOUG	DOUAGN	GAGN	ACCAGN	e_μ(train)	e_μ(test)	A(train)	A(test)
1	.1877	.1768	.0926	.0862	.0452	.0090	.0131	98.65	96.96
2	.1915	.1778	.0983	.0758	.0447	.0098	.0060	98.65	100.00
3	.1847	.1801	.0901	.0752	.0469	.0095	.0082	98.65	100.00
4	.1888	.1804	.0938	.0743	.0467	.0096	.0073	98.65	100.00
5	.1882	.1775	.0908	.0767	.0436	.0097	.0061	98.31	100.00
6	.1827	.1826	.0930	.0746	.0454	.0096	.0072	98.65	100.00
7	.1919	.1767	.0940	.0685	.0480	.0096	.0073	98.98	100.00
8	.1885	.1780	.0961	.0813	.0460	.0088	.0143	98.98	93.93
9	.2001	.1754	.0909	.0711	.0458	.0090	.0134	99.32	93.93
10	.1852	.1789	.0944	.0712	.0446	.0089	.0133	99.32	96.96
mean	.1890	.1785	.0935	.0755	.0457	.0094	.0097	98.82	98.18
95% CI	.0112	.0048	.0057	.0119	.0029	.0009	.0077	.75	5.86

call that ACCENTFOCUS and DOUBLEGROUND were classified as hard in Experiment 12). The other three constraints receive low weights characteristic of soft constraints. As in Experiment 11, DOUBLEALIGN and GROUNDALIGN differ only marginally, they are assigned weights of .0935 and .0755, respectively. The fact that the confidence intervals overlap suggests a tie in constraint rank. The constraint ACCENTALIGN has a low ranking of .0457, indicating that this a violation of ACCENTALIGN has only a small effect on the acceptability of a structure. We arrive at the overall hierarchy $\{\text{ACCENTFOCUS}, \text{DOUBLEGROUND}\} \gg \{\text{DOUBLEALIGN}, \text{GROUNDALIGN}\} \gg \text{ACCENTALIGN}$. Note that this ranking is consistent with the one induced for the data from Experiment 11.

Separate LOT analyses were carried out for each of the five contexts included in Experiment 12, yielding the constraint weights graphed in Figure 7.12 (again some context fail to provide evidence for certain constraints). Recall that we predicted that ACCENTFOCUS and the DOUBLEGROUND are hard constraints and thus should be immune to context-specific reranking. This prediction is clearly borne out with respect to the DOUBLEGROUND, which

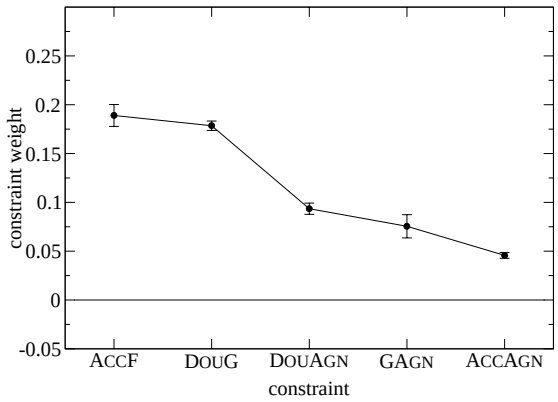

Figure 7.11: Constraint weights for Greek word order data (Experiment 12)

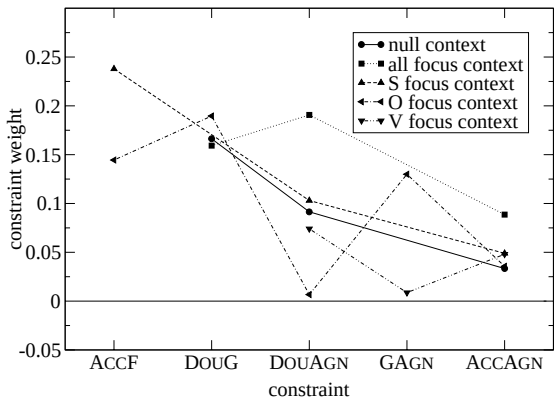

Figure 7.12: Context effects for Greek word order data (Experiment 12)

receives very similar weights in all contexts. The result for ACCENTFOCUS is less clear; its weight is higher in the S context than in the O context. This leaves open the possibility that ACCENTFOCUS is a soft constraint, even though it receives fairly high weights in both contexts.

The constraints DOUBLEALIGN and GROUNDALIGN, however, are clearly soft, as they exhibit large context-specific re-ranking effects. For DOUBLEALIGN, we obtain a high weight in the all focus context, consistent with the modeling study for Experiment 11 and with the observations concerning Experiment 12 reported in Section 4.7.5. Also note that the constraint ACCENTALIGN, which we classified as soft due to its low constraint weight, shows some effects of re-ranking; its weight is highest in the all focus context.

7.6.3. Model Fit and Predictions

Tables 7.21 and 7.22 report the model fit for the two modeling studies for the Greek word order data. For the model based on Experiment 11, we find a low average MSE of .0056, and a perfect accuracy of 100% on the training set (the mean standard deviation for the experimental data was .2908). The model's prediction generalize well to unseen data: the average MSE on the test data is only slightly higher at .0061, while the accuracy is again 100%.

The performance of the model for Experiment 12 is slightly lower than the model for the first experiment. The average MSE on the training data is .0094, corresponding to an accuracy of 98.82% (the mean standard deviation for the experimental data was .2682). The model performs similar on unseen data: the MSE on the test set is only slightly lower at .0097, while the accuracy on the test set is 98.18%. This shows that the model is able to generalize.

7.6.4. Conclusions

The present modeling study dealt with word order preferences in Greek and yielded plausible constraint ranks for the data obtained in Experiment 11 and 12. Like Modeling Study 4, the present study demonstrated how LOT can be used to generate a constraint hierarchy for each of the contexts under investigation, thus allowing us to establish which constraints exhibit context-specific re-ranking (recall that context-specific re-ranking serves as a diagnostic for constraint type).

Note that the LOT models presented in Modeling Studies 1–4 fail to exploit the full potential of LOT, in the sense that they simply constitute an LOT implementation of an underlying factorial design used for the ANOVA in the respective experiments—this factorial design is encoded as an LOT violation profile. The present modeling study goes beyond this: it uses constraint violation profiles that cannot be expressed in terms of a factorial design (see Section 4.6.3.3): the corresponding factorial design would contain empty cells since no linguistic structures exist that fit in these cells (i.e., that exhibit the right combination of constraint violations and constraint satisfactions required by the cell). This demonstrates that an LOT approach is more general than an ANOVA-based approach, it can be applied to candidate sets that do not fit into a factorial design.

7.7. Comparison with Other Analytic Methods

This section raises the question of how the LOT approach compares with other analytic methods such as analysis and of variance (ANOVA) or multiple regression (MR). It could be argued that these methods constitute simpler, more standard alternatives to LOT and its Least Square Estimation scheme.

7.7.1. Analysis of Variance

Analysis of variance was used extensively in Chapters 3 and 4 to establish the validity of a given set of linguistic constraints. While ANOVA is an excellent tool for determining whether a constraint has a significant effect on the acceptability of a given linguistic structure, it provides no straightforward way of quantifying the influence of the constraint. This means that ANOVA is not suitable for estimating constraint weights or ranks, a crucial element of the LOT model.[3]

A second aspect concerns experimental design. It is not always possible to formulate a given theoretical question in terms of a factorial design, as is necessary when conducting an ANOVA. This was illustrated in Modeling Study 5 which presented an LOT model of Experiments 11 and 12. In the case of these experiments, the set of constraints used for analyzing the results failed to match the set of factors used in the experimental design. This has the consequence that the ANOVA results cannot provide direct evidence for the ranking of the constraints. In the general case, we cannot expect a candidate set that is of theoretical relevance to have a violation profile that neatly fits into a factorial design; typically it will contain empty cells, thus preventing a straightforward ANOVA.

This problem is addressed by an LOT approach, which imposes fewer restrictions on the experimental design than a the factorial setup required for an ANOVA.

7.7.2. Multiple Regression

An alternative approach to analyzing gradient data is provided by multiple regression. In contrast to analysis of variance, multiple regression allows us to quantify the influence of a given factor (constraint) on acceptability. Also, MR imposes fewer restrictions on the experimental design than ANOVA (which can be regarded as a special case of MR, see Edwards 1984; Rietveld and van Hout 1993).

Multiple regression assumes that a dependent variable (here acceptability) can be predicted by a linear combination of experimental factors. In this respect, MR is similar to LOT (as was already discussed briefly in Section 6.4.4). We could therefore apply multiple regression to a set of acceptability judgments, with the constraint violations coded as the factors in the regression equation. MR would then return a set of coefficients for these factors, which can be interpreted as constraint weights.

However, the LOT model differs in two important aspects from MR. Firstly, LOT is informed by linguistic theory in a way that MR is not. An LOT data set consists of ranking arguments, not of raw acceptability judgments. This is desirable from a theoretical point of view, as we are only interested in modeling acceptability *differences*, not absolute acceptability judg-

[3] It might be possible, however, to derive constraint weights from measures of effect size, such as the η metric available for the ANOVA (η^2 can be interpreted as the amount of variance accounted for by a given variable). We leave this question for further research. (See Cowart 1997 for some discussion on the interpretation of η in experiments on linguistic judgments.)

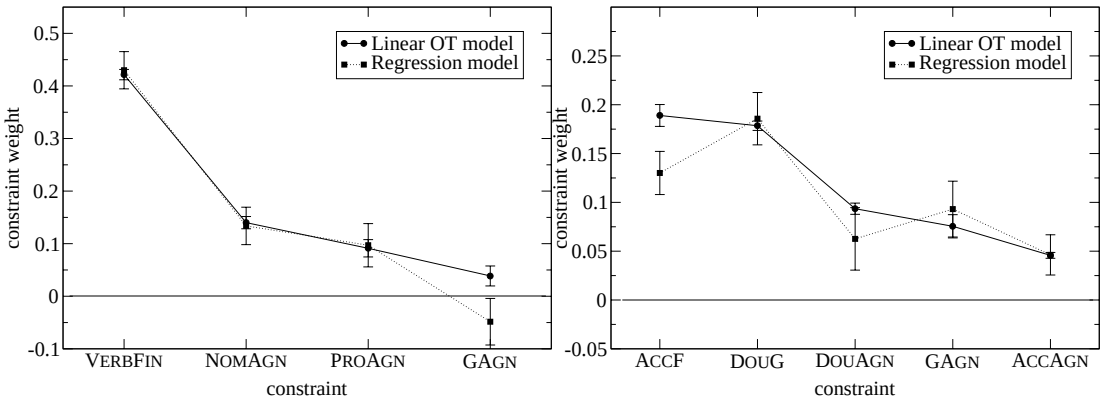

Figure 7.13: Comparison of Linear Optimality Theory and multiple regression models (Experiments 10 and 12)

ments, which can vary from experiment to experiment due to non-linguistic factors. Note also that weight estimation for LOT works on candidate sets, not on the whole set of acceptability judgments. Therefore, an LOT model only accounts for the acceptability differences between structures that compete in an optimality theoretic sense; LOT does not use comparisons across candidate sets (which are not meaningful in LOT). An MR model, on the other hand, tries to fit a model that accounts for *all* acceptability differences in the data set, even across candidate sets.

The second crucial difference between LOT and MR is that LOT is more restrictive, as it is based on specific, theoretically motivated assumptions about how constraints interact. LOT assumes that only constraint *violations* need to be taken into account in computing acceptability; constraint *satisfactions* are not assumed to influence acceptability. This entails that constraint weights in LOT are always positive; a negative weight would entail that a constraint violation increases acceptability. An MR model does not include such restrictions; it allows positive and negative constraint weights, both constraint violations and constraint satisfactions can be included in the model (as in the model for Experiment 10, see below). Therefore, the MR framework is less predictive than the LOT framework, as it imposes less restrictions on the space of possible models it allows. (For a more detailed discussion of the difference between models that allows only positive weights and model that allow both positive and negative weights, see Section 6.4.4).

We will illustrate the differences between the LOT approach and the MR approach by fitting MR models to the data from Experiment 10 and 12. These MR models yield different constraint weights from the ones derived using LOT in Modeling Studies 4 and 5.

For Experiment 10 (see Figure 7.13a), LOT and MR yield approximately the same constraint weights for the three constraints VERBFINAL, NOMALIGN, and PROALIGN. For the context specific constraint GROUNDALIGN, however, LOT yields a weight of .0386, while

MR yields a negative weight of $-.0485$. This is a crucial difference as it means that a violation of this constraints *increases* acceptability in the MR model. Such a counterintuitive result can be explained by the fact that an MR approach attempts to model the overall acceptability pattern in the whole data set, i.e., it compares structures across candidate sets. In particular, this means that candidates with different context specifications are compared, which is not meaningful for context-dependent constraints such as GROUNDALIGN.

For Experiment 12 (see Figure 7.13b) we find that the MR model and the LOT model yield the same weights for some constraints (DOUBLEGROUND, ACCENTALIGN) but different weights for others (ACCENTALIGN, ACCENTFOCUS, GROUNDALIGN). This entails differences in constraint rank; the LOT model ranks ACCENTFOCUS and DOUBLEGROUND equally, while the MR model includes the ranking DOUBLEGROUND $\gg$ ACCENTFOCUS. Note that also the relative ranking of DOUBLEALIGN and GROUNDALIGN differs in the two models. A comparison with the context-specific constraint weights in Figure 7.12 shows that the LOT and the MR model differ specifically on the constraints that show the strongest contextual variation. This indicates that the divergence of the models is caused by the fact that LOT takes candidate set restrictions into account, while MR attempts to fit a model to the whole data set, whether a comparison between a set of structures is theoretically meaningful or not.

7.8. Conclusions

This chapter provided a set of detailed Linear Optimality Theory models of experimental data. These models shared a common methodology: we first established a set of constraints, then we identified a candidate set and its constraint violation profile, and based on this we computed a set of ranking argument. These ranking arguments then served as the input for the Least Square Estimation algorithm, which computed a set of weights for the constraints in question. These weights correspond to linguistically meaningful OT-style constraint ranks. Furthermore, we conducted crossvalidation on the data set to establish if the LOT model is predictive, i.e., if it is able to generalize to unseen data.

Apart from illustrating the general LOT modeling approach, we also presented LOT models for specific linguistic phenomena, based on selected data sets from Chapters 3 and 4:

- A model of extraction from picture NPs was presented that demonstrated how the distinction between soft and hard constraints is modeled in LOT. The model was found to assign low constraint weights to soft constraints and high weights to hard constraints.
- A model of gapping illustrated how LOT can capture the distinction between context-dependent and context-independent constraints. Context-dependent constraints exhibit context-specific re-ranking (changes in constraint weight), while context-independent constraints fail to do so.

- A model of word order in Greek and German based on hypothetical data showed that crosslinguistic variation can be accounted for by crosslinguistic re-ranking in LOT, just as in Standard OT.

Note that all these properties of gradient data (soft and hard constraints, context effects, crosslinguistic effects) do not have to be stipulated, but follow naturally from the fact that constraints are weighted and that constraint violations are cumulative, the two core assumptions that underlie the LOT approach.

Finally, we presented a detailed LOT model of word order variation that drew on results from Experiments 6 and 10–12. This model illustrated how LOT can be used in a realistic study incorporating several sets of experimental results and including data from more than one language. The modeling uncovered new linguistic facts in the form of additional constraint rankings and included a detailed account of the effect of context on word order constraints.

We also compared the LOT approach with standard approaches to data analysis that could be conceived as alternatives to the LOT framework presented in this thesis: analysis of variance and multiple regression. We argued that an approach based on analysis of variance is too limited, as it fails to quantify the size of the effect of constraint violations and puts restrictions on the experimental design. A multiple regression approach, on the other hand, is too general, as it fails to implement the restrictions imposed by an optimality theoretic model of gradience. A multiple regression model allows both positive and negative constraint weights and does not take ranking arguments and constraint sets into account. This was demonstrated by a comparison of a LOT model of word order variation with a multiple regression model on the same data.

Chapter 8

Conclusions

This chapter summarizes the main findings of this thesis and outlines some issues for further research raised by these findings.

8.1. Main Findings

This thesis investigated gradience in grammar, i.e., the fact that some linguistic structures are not fully acceptable or unacceptable, but receive gradient linguistic judgments. The results of this investigation provided a series of experimental, theoretical, and methodological contributions towards the understanding of gradience. The following is a summary of the central findings:

- We conducted a series of experiments that covered all major syntactic modules and investigated representative syntactic phenomena in three languages. These experiments showed that gradience is a systematic, pervasive grammatical phenomenon, and that gradient experimental data can yield insights that are not readily available from intuitive, informal linguistic judgments.

- The experimental results demonstrated that all gradient phenomena share a common set of properties: (a) constraint violations are ranked, i.e., they differ in seriousness; (b) constraint violations are cumulative, i.e., the degree of unacceptability increases with the number of violations.

- The experimental findings also supported a systematic distinction between hard and soft constraints, which can be operationalized using three criteria: (a) soft constraints lead to mild unacceptability when violated, while hard constraint violations trigger serious unacceptability; (b) context effects only occur for soft constraints; hard constraints are immune to contextual variation; (c) the soft/hard distinction is crosslinguistically stable.

- It was shown that magnitude estimation, which so far has only been applied to isolated sentences, can also be used for contextualized acceptability judgments and for coreference judgments. We found that isolated sentences behave like sentences presented in a minimally informative (all focus) context.

- We also demonstrated that magnitude estimation experiments can be carried out over the world wide web; the results obtained this way were highly correlated with those achieved in the laboratory or using questionnaires.

- To account for our experimental results on gradience in grammar, we proposed a model called Linear Optimality Theory. This model, which borrows core concepts from Optimality Theory, is based on the assumptions that constraints are ranked and that constraint violations are cumulative. These assumption can be implemented using weighted constraints and a linear constraint combination scheme.

- Parameter estimation, i.e., the task of determining the constraint weights for a Linear Optimality Theory model, was shown to reduce to the problem of solving a system of linear equations. Standard algorithms with attractive computational properties are available to solve this problem.

- We proved that Standard Optimality Theory is a special case of Linear Optimality Theory where constraint weights are computed in an exponential fashion. Linear Optimality Theory, on the other hand, can be simulated in Standard OT if stratified constraint hierarchies are allowed. This suggests that existing OT-based analyses will carry over to Linear Optimality Theory.

- A series of modeling studies based on our experimental data demonstrated that central properties of gradient data (soft/hard distinction, context effects, crosslinguistic effects) follow from the core assumptions of Linear Optimality Theory, and do not have to be stipulated separately.

Taken together, this thesis contributed an explicit model of gradience in grammar. This model is motivated by extensive experimental studies and is grounded in linguistic theory as it draws on key concepts from Optimality Theory. We demonstrated how detailed, theoretically meaningful linguistic analyses of gradient phenomena can be obtained using this model.

8.2. Issues for Further Research

In this section, we will provide a brief discussion of a number of issues for further research that follow from the findings reported in this thesis.

8.2.1. Further Modeling Studies

Further modeling studies should be carried out to back up the claim that LOT offers a suitable framework for accounting for gradient linguistic data. In particular, it should be demonstrated in more detail that LOT can deal with crosslinguistic variation. (Recall that only a proof of concept study was provided in Chapter 7.)

An obvious starting point for a further investigation of crosslinguistic variation is the data from Experiments 1–3, where we investigated auxiliary selection and impersonal passive formation in unaccusative and unergative verbs. These experiments dealt with two dialects of German, and thus provide a testing ground for an LOT model of crossdialectal variation (which we take to be an instance of crosslinguistic variation). Such an LOT study of unaccusativity and unergativity could be expanded by taking into account the data for Italian collected by Bard et al. (1996) and the data for Dutch provided by Sorace and Vonk (1998). The challenge is to develop a constraint set that is universal for all of these languages, and whose weights can be determined using Least Square Estimation on the judgment data.

8.2.2. Diagnostics for Constraint Type

In Chapter 7 we illustrated two kinds of constraint re-ranking, both of which can serve as diagnostics of the type of a constraint (soft or hard). The first one is context-specific re-ranking, i.e., the fact that the rank of a given constraint varies across contexts. We adopted the hypothesis that context-specific re-ranking only occurs for soft constraints, and this claim was well supported by the experimental data presented in Chapter 4.

As a second diagnostic of constraint type we proposed crosslinguistic re-ranking, i.e., the fact that the rank of a given constraint varies from language to language. We claimed that crosslinguistic re-ranking occurs in both soft and hard constraints, but hypothesized that it does not change the type of a constraint, i.e., that a given constraint will be soft or hard across languages. We provided some evidence for this hypothesis in Experiments 1–3, which dealt with crossdialectal variation in unaccusative and unergative verbs. Experiments 10–12 provided further support by demonstrating that the Information Structure constraint GROUNDALIGN is crosslinguistically soft. Also the proof of concept study in Section 7.4 suggests that constraints like NOMALIGN and DATALIGN that regulate complement order are crosslinguistically soft, while constraints like VERBFINAL and VERBINITIAL that regulate verb position are crosslinguistically hard.

However, the experimental support for crosslinguistic re-ranking is weaker than the one for context-specific re-ranking. Experimental studies explicitly designed to test the crosslinguistic re-ranking hypothesis are necessary and should be the subject of further research.

Linear Optimality Theory also generates a more general hypothesis regarding the

soft/hard distinction, viz., that all diagnostics of the type of a constraint can be reduced to constraint re-ranking. It seems plausible to assume that other diagnostics of the type of a constraint exist in addition to contextual and crosslinguistic effects. The LOT model should be able to accommodate these diagnostics in terms of constraint re-ranking. Preliminary evidence for this prediction comes from language development, and will be discussed in the next section.

8.2.3. Gradience and Language Development

In Section 6.4.2.6 we provided some speculations on the role of gradient data in language development, and argued that the LOT model might be suitable for accounting for gradient data that occurs in first and second language acquisition and in language attrition.

It has been argued in the second language acquisition literature (Papp 2000; Prévost and White 2000; Robertson 2000) that certain parts of the second language grammar exhibit optionality, i.e., the grammar admits more than one structure as a realization of a given input. Typically, such optional structures are not equally acceptable, but differ in their degree of acceptability. The relative acceptability of optional structures changes during second language acquisition and eventually approximates the acceptability pattern in the native grammar. Sorace (1993a,b) demonstrates this effect for auxiliary selection in French and Italian.

Optionality and gradience have also been reported for first language acquisition. Adult and non-adult structures co-exist in the grammar of the language learner, with the non-adult structures gradually disappearing during language development. A relevant example are non-finite verb structures that alternate with finite ones in early child grammar; this phenomenon has been modeled using OT by Legendre, Vainikka, Todorova, and Hagstrom (1998), who provide an account that predicts the gradual reduction of infinite structures during language acquisition (for other OT-based accounts of first language acquisition see Boersma and Levelt 1999; Lee 1998).

A third aspect of language development is language attrition. Attrition refers to changes in the native language grammar that occur in second language users. According to Sorace (1998, 1999), attrition can be characterized as the emergence of gradience (i.e., optionality) in the native language caused by sustained exposure to a second language. A relevant example is pro-drop in Italian. Pro-drop of referential subjects is obligatory in Italian if the subject is coreferential with a topic antecedent; otherwise the subject is overtly expressed. This constraint is subject to attrition in native speakers of Italian that are exposed to English as a second language. These speakers overproduce overt subjects even when coreferentiality with a topic obtains. This suggests that pro-drop is optional in the grammars of these speakers: null subjects are produced in appropriate contexts, but overt subjects are optionally produced in inappropriate contexts. It is interesting that the reverse effect is not observed, i.e., native speakers of English under sustained exposure to Italian as a second language do not exhibit optional pro-drop in English. This phenomenon lends itself to an optimality theoretic treatment (Sorace

1998, 1999).

Two important questions arise concerning the role of LOT in language development. The first one concerns the formalization of developmental optionality effects. In an LOT setting, two structures are optional if they share the same candidate set and receive similar harmony scores. The relative grammaticality of optional structures depends on the weights of the constraints they violate. This means that developmental changes in the relative acceptability of optional structures can be modeled in LOT as constraint re-ranking. In such a setting, the weights of certain constraints change gradually in the course of language development.

The second point concerns the observation that developmental optionality seems to be limited to certain types of constraints. Optionality in auxiliary selection in French and Italian, for instance, seems to concern only peripheral verb classes, while core verb classes exhibit a binary auxiliary selection behavior (Sorace 1993a,b). Another point in case is the attrition of the pro-drop constraint, described above (Sorace 1998, 1999). A hypothesis for future research is that soft constraints are subject to developmental optionality, while hard constraints are immune to such developmental effects. This would mean that *developmental re-ranking* can serve as a further diagnostic of the hard/soft dichotomy, in addition to crosslinguistic re-raking and contextual re-ranking.

8.2.4. Explaining the Soft/Hard Dichotomy

A central claim of this thesis was that a given linguistic constraints can be classified as either soft or hard, based on its behavior with respect to gradient data. This claim led to diagnostics for the type of a constraint, including violation strength, context effects, and crosslinguistic variation.

Throughout the thesis, we took a purely formal view of the soft/hard dichotomy. We have not attempted to explain *why* constraints come in two types. This is a key area for further research on gradience; in the following, we will provide a number of speculations on future directions.

One hypothesis that could be pursued is that hard constraints are structural (i.e., syntactic) in nature, while soft constraint are non-structural (e.g., semantic or pragmatic). This seems to pan our fairly well with respect to the constraints investigated in Experiments 1–12. For instance, the soft constraints REFERENTIALITY, DEFINITENESS, and VERBCLASS in the extraction and binding data in Experiments 4, 5, and 9 are clearly semantic or pragmatic in nature. The hard constraints AGREEMENT, INVERSION, RESUMPTIVE, and INTERVENE, on the other hand, are clearly structural.

A similar observation can be made with respect to the word order data in Experiments 6 and 10–12. The soft constraints NOMALIGN and DATALIGN make reference to thematic roles and thus can be argued to have a semantic component. The constraint PROALIGN deals with pronominalization, i.e., it is clearly semantic. PROALIGN was classified as a soft constraint

based on its experimental behavior. The hard constraints VERBINITIAL and VERBFINAL, on the other hand, regulate verb position, and therefore belong to the domain of syntax.

The hypothesis that hard constraints are structural makes interesting predictions for the constraints ACCENTFOCUS and DOUBLEGROUND (see Experiment 12). Both constraints were classified as hard constraints as they induce strong unacceptability and fail to show contextual variation. This is expected under an account of Information Structure that assumes that focus and ground are part of the structural component of the grammar (e.g., Steedman 1991).

Note that the hypothesis that soft constraints are semantic/pragmatic, while hard ones are syntactic, is also compatible with Sorace's (1999) claim that interpretable features (in Chomsky's (1995) terms) are subject to language attrition, while non-interpretable features are immune to attrition (see the discussion of gradience and language development in Section 8.2.3).

8.2.5. Ranking Algorithms

In Section 6.3.8 we showed that Standard OT is a special case of Linear Optimality Theory, i.e., a Standard OT grammar can be simulated in LOT by assigning the weights in an exponential fashion. It remains an open question if this result entails that the LOT learning algorithm that we proposed in Section 6.3.5 can be applied to learn Standard OT grammars. A positive answer to this questions would mean that a straightforward, well-understood learning algorithm for Standard OT is available in the form of Least Square Estimation. This would be a remarkable achievement, given that only partial solution to the OT learning problem are available in the existing literature (Boersma 1998; Tesar and Smolensky 1998). Note that Least Square Estimation offers attractive complexity properties; its data complexity is linear, and its time complexity is polynomial (see Section 6.3.6).

The main obstacle to applying Least Square Estimation to Standard OT concerns the fact that Least Square Estimation makes use of gradient ranking arguments, instead of the binary ranking arguments that Standard OT learning schemes use. A gradient ranking argument consists of two structures for which the violation profiles and the acceptability difference is known; for a binary ranking argument, we only need to know the violation profiles and which structure is the winner; no information about relative acceptability is required. It is an open question if binary ranking arguments as they are available in Standard OT can be enriched so that they can serve as input for Least Square Estimation. One possibility would be to set the acceptability difference between the winner and the loser to a constant amount. However, this might not be sufficient for the LSE scheme to compute meaningful constraint weights for the constraint set.

8.2.6. Computing Significant Constraint Weights

One problem that remains is that LSE does not specify if a constraint has a significant effect on acceptability (in contrast to, for instance, linear regression). We know that a constraint whose weight is zero (or negative) does not influence acceptability. However, there are cases where a constraint receives a low weight (examples are the constraints REFERENTIALITY, DEFINITENESS, and VERBCLASS in the modeling study in Section 7.2), but nevertheless has a significant influence on acceptability (as determined by the ANOVA in Experiment 4). Ideally, we should be able to distinguish such a situation from cases where a constraint receives a low weight because it has no significant effect on acceptability. It seems possible that standard techniques from linear regression can be applied to determine the significance of constraint weights in an LOT model.

Another question concerns the comparison of the weights of two constraints; ideally, we would like to have a test to determine if two constraint weights are significantly different from each other. Using crossvalidation provides confidence intervals for the weight estimates, which goes some way towards solving this problem. It can be hypothesized that two constraint weights are significantly different if their confidence intervals fail to overlap.

A third problem concerns the correlation of constraints. If two constraints are highly correlated (i.e., if they are met and violated in the same structures in the candidate set), the LSE method will favor the constraint that accounts best for the data, and assign it a high weight, whereas the constraint that is correlated with it will receive only a low weight, even if it is only slightly worse in accounting for the data. (This phenomenon—collinearity—is well-known from linear regression.) Note however, that such a situation seems to be rare; however, it occurred in the data of Experiment 11, where the constraints DOUBLEALIGN and VERBALIGN were correlated (see Section 7.6).

Appendix A

Instructions

We only give the instructions for Experiment 4; the other experiments used the same type of instructions. For experiments that presented contextualized stimuli (Experiments 7–12), the instructions were modified to take context into account. A modified version of the instructions was also used for Experiments 5 and 14, where subjects were asked to judge coreference instead of to acceptability. The instructions were translated into German for Experiments 1–3, 6, and 10, and translated into Greek for Experiment 11 and 12.

Experiment on Sentence Judgments

Thanks for taking part in this experiment!

Please read the instructions carefully before starting. Do not hesitate to contact the experimenter in case you have any questions or comments concerning this experiment.

This experiment requires a Java compatible web browser and Java has to be enabled.

Depending on the hardware and browser you use, and on your net connection, the execution of the experiment may be slow at times.

Personal Details

As part of this experiment, we have to collect a small amount of personal information, which we ask you to enter in the Personal Details window below. *This information will be treated confidential, and will not be made available to a third party. None of the responses collected in this experiment will be associated with your name in any way.* If you have any questions about this practice, please contact the experimenter.

Please be careful to fill in the Personal Details questionnaire correctly, as otherwise we will have to discard your responses.

We ask you to supply the following information:

- your name and email address;
- your age and sex;
- whether you are right or left handed (based on the hand you prefer to use for writing);
- the academic subject you study or have studied (or your current occupation in case you haven't attended university);
- under "Region", please specify the place (city, region/state/province, country) where you have learned your first language.

Instructions

Part 1: Judging Line Length

Before doing the main part of the experiment, you will do a short task involving judging line length. A series of lines of different length will be presented on the screen. Your task is to estimate how long they seem by assigning numbers to them. You are supposed to make your estimates relative to the first line you will see, your *reference* line. Give it any number that seems appropriate to you, bearing in mind that some of the lines will be longer than the reference and some will be shorter.

After you have judged the reference line, assign a number to each following line so that it represents how long the line is in proportion to the reference. The longer it is compared to the reference, the larger the number you will use; the shorter it is compared to the reference, the smaller the number you will use. So if you feel that a line is twice as long as the reference, give it a number twice the reference number; if it's a third as long, provide a number a third as big as the reference.

So, if the reference is this line, you might give it the number 10:

If you have to judge this line, you might assign it 17:

And this one might be 2.5:

There is no limit to the range of numbers you may use. You may use whole numbers or decimals. If you assigned the reference line the number 1, you might want to call the last one 0.25. Just try to make each number match the length of the line as you see it.

Parts 2 and 3: Judging Sentences

In Part 1 of the experiment you used numbers to estimate the length of lines on the screen. In Parts 2 and 3 you will use numbers to judge the acceptability of some English sentences in the same way.

You will see a series of sentences presented one at a time on the screen. Each sentence is different. Some will seem perfectly OK to you, but others will not. Your task is to judge how good or bad each sentence is by assigning a number to it.

As with the lines in Part 1, you will first see a *reference* sentence, and you can use any number that seems appropriate to you for this reference. For each sentence after the reference, you will assign a number to show how good or bad that sentence is in proportion to the reference sentence.

For example, if the reference sentence was:

(1) The dog the bone ate.

you would probably give it a rather low number. (You are free to decide what "low" or "high" means in this context.) If the next example:

(2) The dog devoured yesterday the bone.

seemed 10 times better than the reference, you'd give it a number 10 times the number you gave to the reference. If it seemed half as good as the reference, you'd give it a number half the number you gave to the reference.

You can use any range of positive numbers that you like, including decimal numbers. *There is no upper or lower limit to the numbers you can use,* except that you cannot use zero or negative numbers. *Try to use a wide range of numbers and to distinguish as many degrees of acceptability as possible.*

There are no "correct" answers, so whatever seems right to you is a valid response. We are interested in your first impressions, so please don't take too much time to think about any one sentence: try to make up your mind quickly, spending less than 10 seconds on sentence.

Procedure

First please fill in the Personal Details questionnaire as explained above, and then press the Start button.

The experiment will consist of the following 3 parts:

- Training session: judging 6 lines
- Practice session: judging 6 sentences
- Experiment session: judging 32 sentences

In each part you will see the reference item in the experiment window. Please enter your reference number and then press the Continue button. Now the test items will appear one after the other in the experiment window. Please type your judgment in the box below each item.

The experiment will take 5 to 10 minutes. After the experiment is completed you will receive an email confirmation of your participation.

Please keep in mind:

- Use any number you like for the reference sentence.
- Judge each sentence in proportion to the reference.
- Use any positive numbers which you think are appropriate.
- Use high numbers for “good” sentences, low numbers for “bad” sentences and intermediate numbers for sentences which are intermediate in acceptability.
- Try to use a wide range of numbers and to distinguish as many degrees of acceptability as possible.
- Try to make up your mind quickly, base your judgments on your first impressions.

Appendix B

Materials

B.1. Experiment 1

The stimuli for the auxiliary selection experiment are listed in (B.1)–(B.8) for each verb class. The stimuli for the impersonal passive experiment were derived from the stimuli in (B.1)–(B.8) by replacing the subject and the auxiliary with *es wurde* “it was”.

(B.1) **Change of Location**

a. Die Bergsteigerin hat/ist vorsichtig aufgestiegen.
the climber has/is carefully ascended
“The climber ascended carefully”

b. Der Gefangene hat/ist schnell entkommen.
“The prisoner escaped quickly.”

c. Die Besucherin hat/ist sofort zurückgekommen.
“The visitor returned immediately.”

d. Der Gast hat/ist pünktlich angekommen.
“The guest arrived on time.”

e. Die Touristin hat/ist überstürzt abgereist.
“The tourist departed hastily.”

f. Der Verbrecher hat/ist flink geflüchtet.
“The criminal fled quickly.”

g. Die Besucherin hat/ist hastig weggegangen.
“The visitor left hastily.”

h. Der Soldat hat/ist vorsichtig vorgerückt.
“The soldier advanced carefully.”

(B.2) **Change of State**

a. Die Bewerberin hat/ist pünktlich erschienen.
the applicant hat/is on time appeared
“The applicant appeared on time.”

b. Der Prüfling hat/ist langsam erblasst.
"The examinee slowly became pale."
c. Die Angestellte hat/ist schnell nervös geworden.
The employee quickly became nervous.
d. Der Großvater hat/ist unerwartet verstorben.
"The grandfather died unexpectedly."
e. Das Mädchen hat/ist langsam errötet.
"The girl blushed slowly."
f. Der Badende hat/ist langsam erkaltet.
"The bather slowly became cold."
g. Das Kind hat/ist schnell gewachsen.
"The child grew quickly."
h. Das Kind hat/ist bald verschwunden.
The child disappeared quickly."

(B.3) **Continuation of State**

a. Die Bettlerin hat/ist elend dahinvegetiert
the beggar has/is miserably vegetated
"The beggar vegetated miserably."
b. Der Einsiedler hat/ist zäh überdauert.
"The hermit survived tenaciously."
c. Die Soldatin hat/ist tapfer ausgehalten.
"The soldier endured courageously."
d. Der Flüchtling hat/ist elend weiterexistiert.
"The refugee continued to exist miserably."
e. Die Kranke hat/ist glücklich weitergelebt.
"The sick person continued to live happily."
f. Der Patient hat/ist glücklich überlebt.
"The patient survived happily."
g. Die Wartende hat/ist regungslos verharrt.
"The waiting person tarried motionlessly."
h. Der Wanderer hat/ist kurz verweilt.
"The hiker stayed for a short while."

(B.4) **Existence of State (Positional)**

a. Die Polizistin hat/ist ratlos herumgestanden.
the policewoman has/is cluelessly stand about
"The policewoman stood about cluelessly."
b. Der Jugendliche hat/ist gelangweilt herumgehangen.
"The teenager hung about lackadaisically."

c. Die Betende hat/ist würdevoll gekniet.
"The praying person keeled with dignity."

d. Der Gefangene hat/ist angstvoll gekauert.
"The prisoner crouched fearfully."

e. Die Artistin hat/ist hoch oben gebaumelt.
"The trapez artist dangled high up."

f. Der Drachenflieger hat/ist hoch oben geschwebt.
"The paraglider hovered high up."

g. Das Schulkind hat/ist gebückt gesessen.
"The pupil sat crookedly."

h. Das Kind hat/ist bewegungslos gehockt.
"The child squatted motionlessly."

(B.5) **Controlled, Non-motion**

a. Die Lehrerin hat/ist dauernd geredet.
the teacher has/is continously talked
"The teacher talked continously."

b. Der Professor hat/ist würdevoll doziert.
"The professor lectured in a dignified way."

c. Die Nachbarin hat/ist aufgeregt geplaudert.
"The neighbor chat excitedly."

d. Der Mann hat/ist angstvoll gewartet.
"The man waited fearfully."

e. Die Frau hat/ist schwer gearbeitet.
"The woman worked hard."

f. Der Nachbar hat/ist lange telefoniert.
"The neighbor phone for a long time."

g. Das Kind hat/ist widerwillig nachgegeben.
"The child gave in reluctantly."

h. Das Kind hat/ist begeistert mitgespielt.
"The kind participated enthusiastically."

(B.6) **Controlled, Motion**

a. Die Frau hat/ist schnell geschwommen.
the woman has/is quickly swam
"The woman swam quickly."

b. Der Urlauber hat/ist zügig gewandert.
"The holiday maker hiked briskly."

c. Die Nachbarin hat/ist langsam geschlurft.
"The neighbor shuffled slowly."

d. Der Mann hat/ist schnell gerannt.
"The man ran quickly."
e. Die Tänzerin hat/ist langsam getanzt
"The dancer danced slowly."
f. Der Urlauber hat/ist langsam geklettert
"The holiday maker climbed slowly."
g. Das Kind hat/ist langsam gekrochen
"The child crept slowly."
h. Das Kind hat/ist schnell gehüpft
"The child bounced quickly."

(B.7) **Uncontrolled, Involuntary Reaction**

a. Die Frau hat/ist etwas getorkelt.
the woman has is slightly tottered
"The woman tottered slightly."
b. Der Junge hat/ist stark getaumelt
"The boy staggered heavily."
c. Das Mädchen hat/ist unsicher gewackelt
"The girl waggled insecurely ."
d. Der Mann hat/ist stark geschwankt
"The man wobbled heavily."
e. Das Mädchen hat/ist angstvoll geschaudert
"The girl shuddered with fear."
f. Der Nachbar hat/ist erregt gebebt.
"The neighbor trembled with excitement."
g. Die Frau hat/ist angstvoll gezittert
"The woman jittered with fear."
h. Der Junge hat/ist stark geschlottert
"The boy shivered heavily."

(B.8) **Uncontrolled, Emission**

a. Der Zug hat/ist laut gerumpelt.
the train has/is noisily rattled
"The train rattled noisily."
b. Das Fahrrad hat/ist leise geklappert
"The bike rattled gently."
c. Die U-Bahn hat/ist etwas gebrummt
"The subway buzzed a bit."
d. Das Dreirad hat/ist leise gequietscht
"The tricycle squeaked gently."

e. Das Motorrad hat/ist laut gerattert
   “The motorbike clattered noisily.”
f. Das Schiff hat/ist laut getuckert
   “The ship tapped noisily.”
g. Der Aufzug hat/ist leise gesurrt
   “The elevator whirred gently.”
h. Die Bergbahn hat/ist ein wenig geächzt
   “The funicular moaned a bit.”

The sentence in (B.9) was presented as the modulus. It contains a violation of a word order constraint, as the dative NP *dem Dieb* precedes the nominative NP *der Nachbar.*

(B.9) Daniela gibt zu, dass dem Dieb der Nachbar das Auto leiht.
Daniela admits that the thief the neighbor the car lends
“Daniela admits that the neighbor will lend the car to the thief.”

B.2. Experiment 2

The stimuli for the auxiliary selection experiment are listed in (B.10)–(B.17) for each verb class. The stimuli for the impersonal passive experiment were derived from the stimuli in (B.10)–(B.17) by replacing the subject and the auxiliary with *es wurde* “it was”.

(B.10) **Change of Location, Animate**

a. Die Bergsteigerin hat/ist vorsichtig aufgestiegen.
   the climber has/is carefully ascended
   “The climber ascended carefully”
b. Der Gefangene hat/ist schnell entkommen.
   “The prisoner escaped quickly.”
c. Die Besucherin hat/ist sofort zurückgekommen.
   “The visitor returned immediately.”
d. Der Gast hat/ist pünktlich angekommen.
   “The guest arrived on time.”
e. Die Touristin hat/ist überstürzt abgereist.
   “The tourist departed hastily.”
f. Der Verbrecher hat/ist flink geflüchtet.
   “The criminal fled quickly.”
g. Die Besucherin hat/ist hastig weggegangen.
   “The visitor left hastily.”
h. Der Soldat hat/ist vorsichtig vorgerückt.
   “The soldier advanced carefully.”

(B.11) **Change of State, no Prefix, Inanimate**

a. Die Dose hat/ist sofort gerostet.
the can has/is immediately rusted
"The can rusted immediately."

b. Der Baumstumpf hat/ist langsam gemodert.
"The snag rotted slowly."

c. Der Apfel hat/ist bald gefault.
"The appled rotted soon."

d. Der Käse hat/ist schnell geschimmelt.
"The cheese moldered quickly."

e. Die Rose hat/ist sofort geblüht.
"The rose bloomed immediately."

f. Die Blume hat/ist langsam gewelkt.
"The flower wilted slowly."

g. Der Setzling hat/ist bald gekeimt.
"The seedling germinated soon."

h. Der Gewinn hat/ist schnell gewachsen.
"The profit grew quickly."

(B.12) **Change of State, Prefix, Inanimate**

a. Die Dose hat/ist sofort verrostet.
the can has/is immediately rusted
"The can rusted immediately."

b. Der Baumstumpf hat/ist langsam vermodert.
"The snag rotted slowly."

c. Der Apfel hat/ist bald verfault.
"he appled rotted soon."

d. Der Käse hat/ist schnell verschimmelt.
"The cheese moldered quickly."

e. Die Rose hat/ist sofort erblüht.
"The rose bloomed immediately."

f. Die Blume hat/ist langsam verwelkt.
"The flower wilted slowly."

g. Der Setzling hat/ist bald aufgekeimt.
"The seedling germinated soon."

h. Der Gewinn hat/ist schnell angewachsen.
"he profit grew quickly."

(B.13) **Continuation of State, Inanimate**

a. Die Vorstellung hat/ist lange gedauert.
the performance has/is long lasted
"The performance lasted a long time."

b. Der Regen hat/ist nur kurz angedauert.
"The rain lasted only a short while."

c. Die Auseinandersetzung hat/ist lange fortgedauert.
"The dispute continued for a long time."

d. Das Brot hat/ist lange gehalten.
"The bread lasted for a long time."

e. Der Streit hat/ist nur kurz angehalten.
"The quarrel continued only for a short while."

f. Der Verdienst hat/ist lange gereicht.
"The wage suffice for a long time."

g. Das Wasser hat/ist gerade ausgereicht.
"The water sufficed just about."

h. Das Essen hat/ist gerade genügt.
"The food sufficed just about."

(B.14) **Existence of State (Positional), Animate**

a. Die Täterin hat/ist betreten dagestanden.
the offender is/has sheepishly stood there
"The offender stood there sheepishly."

b. Die Polizistin hat/ist ratlos herumgestanden.
"The policewoman stood about cluelessly."

c. Der Jugendliche hat/ist gelangweilt herumgehangen.
"The teenager hung about lackadaisically."

d. Die Artistin hat/ist hoch oben gebaumelt.
"The trapez artist dangle high up."

e. Der Schläfer hat/ist bequem gelegen.
"The sleeping person lay comfortably."

f. Der Junge hat/ist faul herumgelegen.
"The boy lay about lazily."

g. Die Patientin hat/ist ruhig dagelegen.
"The patient lay there quietly."

h. Der Drachenflieger hat/ist hoch oben geschwebt.
"The paraglider hovered high up."

(B.15) **Existence of State (Positional), Inanimate**

a. Der Korb hat/ist unbeachtet dagestanden.
the basket is/has unnoticed stood there
"The basked stood there unnoticed."

b. Der Stuhl hat/ist nutzlos herumgestanden.
"The chair stood about uselessly."

c. Das Kleid hat/ist unbeachtet herumgehangen.
"The dress hung about unnoticed."

d. Die Fahne hat/ist hoch oben gebaumelt.
"The flag dangled high up."

e. Der Ball hat/ist weit weg gelegen.
"The ball lay far away."

f. Das Geld hat/ist achtlos herumgelegen.
"The money lay about carelessly."

g. Das Buch hat/ist nutzlos dagelegen.
"The book lay there uselessly."

h. Der Drachen hat/ist hoch oben geschwebt.
"The kite hovered high up."

(B.16) **Controlled, Non-motion, Animate**

a. Die Lehrerin hat/ist dauernd geredet.
the teacher has/is continously talked
"The teacher talked continously."

b. Der Professor hat/ist würdevoll doziert.
"The professor lectured in a dignified way."

c. Die Nachbarin hat/ist aufgeregt geplaudert.
"The neighbor chat excitedly."

d. Der Mann hat/ist angstvoll gewartet.
"The man waited fearfully."

e. Die Frau hat/ist schwer gearbeitet.
"The woman worked hard."

f. Der Nachbar hat/ist lange telefoniert.
"The neighbor phone for a long time."

g. Das Kind hat/ist widerwillig nachgegeben.
"The child gave in reluctantly."

h. Das Kind hat/ist begeistert mitgespielt.
"The kind participated enthusiastically."

(B.17) **Uncontrolled, Involuntary Reaction, Non-motion, Animate**

a. Das Mädchen hat/ist angstvoll geschaudert.
The girl has/is with fear shuddered
"The girl shuddered with fear."

b. Der Nachbar hat/ist erregt gebebt.
"The neighbor trembled with excitement."

c. Die Frau hat/ist angstvoll gezittert.
"The woman jittered with fear."

d. Der Junge hat/ist stark geschlottert.
"The boy shivered heavily."

e. Die Patientin hat/ist plötzlich gezuckt.
"The patient convulsed suddenly."

f. Der Sportler hat/ist stark geschwitzt.
"The sportsman sweated heavily."

g. Die Schülerin hat/ist häufig gegähnt.
"The pupil yawned frequently."

h. Der Fußballer hat/ist stark gekeucht.
"The football player panted heavily."

The sentence in (B.18) was presented as the modulus. It contains a violation of a word order constraint, as the accusative NP *den Bericht* precedes the nominative NP *der Chef.*

(B.18) Ich behaupte, dass den Bericht der Chef im Büro liest.
I claim that the report the boss in-the office reads
"I claim that the boss will read the report in the office."

B.3. Experiment 3

The stimuli for each of the verb classes are given in (B.19)–(B.22). There are two forms for each stimulus, containing a telic and atelic adverbials, or positional and durational adverbials (for the control condition).

(B.19) **Continuation of State**

a. Die Bettlerin hat/ist im Elendsviertel dahinvegetiert.
the beggar has/is in the slum vegetated
"The beggar vegetated in the slum."

b. Die Bettlerin hat/ist lange dahinvegetiert.
"The beggar vegetated for a long time."

c. Der Einsiedler hat/ist in der Wüste überdauert.
"The hermit survived in the desert."

d. Der Einsiedler hat/ist jahrelang überdauert.
"The hermit survived for many years."
e. Die Soldatin hat/ist im Schützengraben ausgehalten.
"The soldier endured in the trench."
f. Die Soldatin hat/ist tagelang ausgehalten.
"The soldier endured for days."
g. Der Flüchtling hat/ist in der Hütte weiterexistiert.
"The refugree continued to exist in the hut."
h. Der Flüchtling hat/ist lange Zeit weiterexistiert.
"The refugree continued to exist for a long time."
i. Die Kranke hat/ist im Rollstuhl weitergelebt.
"The sick person continued to live in the wheelchair."
j. Die Kranke hat/ist jahrelang weitergelebt.
"The sick person continued to live for years."
k. Der Patient hat/ist im Krankenhaus überlebt.
"The patient survived in the hospital."
l. Der Patient hat/ist lange überlebt.
"The patient survived for a long time."
m. Die Wartende hat/ist auf der Bank verharrt.
"The waiting person tarried on the bench."
n. Die Wartende hat/ist minutenlang verharrt.
"The waiting person tarried for minutes."
o. Der Wanderer hat/ist auf dem Rastplatz verweilt.
"The hiker stayed at the resting place."
p. Der Wanderer hat/ist eine lange Zeit verweilt.
"The hiker stayed for a long time."

(B.20) **Existence of State (Positional)**

a. Die Polizistin hat/ist am Tatort herumgestanden.
the policewoman has/is on the crime scene stand about
"The policewoman stood about at the crime scene."
b. Die Polizistin hat/ist stundenlang herumgestanden.
"The policewoman stood about for hours."
c. Der Jugendliche hat/ist in der Kneipe herumgehangen.
"The teenager hung about in the bar."
d. Der Jugendliche hat/ist tagelang herumgehangen.
"The teenager hung about for days."
e. Die Betende hat/ist auf dem Beichtstuhl gekniet.
"The praying person keeled in the confessional."

f. Die Betende hat/ist stundenlang gekniet.
“The praying person keeled for hours.”
g. Der Gefangene hat/ist auf dem Boden gekauert.
“The prisoner crouched on the floor.”
h. Der Gefangene hat/ist stundenlang gekauert.
“The prisoner crouched for hours.”
i. Die Artistin hat/ist am Trapez gebaumelt.
“The trapez artist dangled at the trapez.”
j. Die Artistin hat/ist lange gebaumelt.
“The trapez artist dangled for a long time.”
k. Der Drachenflieger hat/ist über den Bäumen geschwebt.
“The paraglider hovered above the trees.”
l. Der Drachenflieger hat/ist lange Zeit geschwebt.
“The paraglider hovered for a long time.”
m. Das Schulkind hat/ist auf dem Stuhl gesessen.
“The pupil sat on the chair.”
n. Das Schulkind hat/ist den ganzen Tag gesessen.
“The pupil sat all day.”
o. Das Kind hat/ist auf dem Boden gehockt.
“The child squatted on the floor.”
p. Das Kind hat/ist die ganze Zeit gehockt.
“The child squatted all the time.”

(B.21) **Controlled, Notion**

a. Die Frau hat/ist ans Ufer geschwommen.
the woman has/is to the shore swam
“The woman swam to the shore.”
b. Die Frau hat/ist im Fluss geschwommen.
“he woman swam in the lake.”
c. Der Urlauber hat/ist zum Gipfel gewandert.
“The holiday maker hiked to the summit.”
d. Der Urlauber hat/ist in den Alpen gewandert.
“The holiday maker hiked in the alps.”
e. Die Nachbarin hat/ist zur Tür geschlurft.
“The neighbor shuffled to the door.”
f. Die Nachbarin hat/ist im Zimmer geschlurft.
“The neighbor shuffled in the room.”
g. Der Mann hat/ist zur Tür gerannt.
“The man ran to the door.”

h. Der Mann hat/ist im Wald gerannt.
"The man ran in the forest."
i. Die Tänzerin hat/ist in den Saal getanzt.
"The dancer danced into the room."
j. Die Tänzerin hat/ist im Saal getanzt.
"The dancer danced in the room."
k. Der Urlauber hat/ist auf den Gipfel geklettert.
"The holiday maker climbed to the summit."
l. Der Urlauber hat/ist im Gebirge geklettert.
"The holiday maker climbed in the mountains."
m. Das Kind hat/ist zur Tür gekrochen.
"The child crept to the door."
n. Das Kind hat/ist auf dem Boden gekrochen.
"The child crept on the floor."
o. Das Kind hat/ist ins Zimmer gehüpft.
"The child bounced into the room."
p. Das Kind hat/ist auf dem Trampolin gehüpft.
"The child bounced on the trampoline."

(B.22) **Uncontrolled, Emission**

a. Der Zug hat/ist in den Bahnhof gerumpelt.
the train has/is in the station rattled
"The train rattled into the station."
b. Der Zug hat/ist im Bahnhof gerumpelt.
"The rain rattled in the station."
c. Das Fahrrad hat/ist durch die Straße geklappert.
"The bike rattled along the street."
d. Das Fahrrad hat/ist auf der Straße geklappert.
"The bike rattled in the street."
e. Die U-Bahn hat/ist in die Haltestelle gebrummt.
"The subway buzzed into the station."
f. Die U-Bahn hat/ist in der Haltestelle gebrummt.
"The subway buzzed in the station."
g. Das Dreirad hat/ist über den Spielplatz gequietscht.
"The tricycle squeaked across the playground."
h. Das Dreirad hat/ist auf dem Spielplatz gequietscht.
"The tricycle squeaked in the playground."
i. Das Motorrad hat/ist durch die Straße gerattert.
"The motorbike clattered along the street."

j. Das Motorrad hat/ist auf der Straße gerattert.
"The motorbike clattered in the street."
k. Das Schiff hat/ist in den Hafen getuckert.
"The ship tapped into the harbor."
l. Das Schiff hat/ist im Hafen getuckert.
"The ship tapped in the harbor."
m. Der Aufzug hat/ist in den vierten Stock gesurrt.
"The elevator whirred to the fourth floor."
n. Der Aufzug hat/ist im vierten Stock gesurrt.
"The elevator whirred at the fourth floor."
o. Die Bergbahn hat/ist ins Tal geächzt.
"The funicular moaned into the valley."
p. Die Bergbahn hat/ist im Tal geächzt.
"The funicular moaned in the valley."

This experiment used the same modulus as Experiment 1 (see (B.9)), prefixed by a neutral (all focus) context sentence:

(B.23) Was gibt's neues? Daniela gibt zu, dass dem Dieb der Nachbar das Auto leiht.
what is-there new Daniela admits that the thief the neighbor the car lends
"What's new? Daniela admits that the neighbor will lend the car to the thief."

B.4. Experiment 4

The examples in (B.24) and (B.25) illustrate how the stimuli for Experiment 4 were constructed for the subexperiments on hard and soft constraint violations. (B.26) lists the lexicalizations, which were the same for both subexperiments.

(B.24) **Soft Constraint Violations**
a. Which model has Mary taken a photograph of?
b. Which model has Mary taken the photograph of?
c. How many models has Mary taken a photograph of?
d. How many models has Mary taken the photograph of?
e. Which opponent has Catherine destroyed a photograph of?
f. Which opponent has Catherine destroyed the photograph of?
g. How many opponents has Catherine destroyed a photograph of?
h. How many opponents has Catherine destroyed the photograph of?

(B.25) **Hard Constraint Violations**
a. Which model has Thomas taken a photograph of?
b. Which model have Thomas taken a photograph of?

c. Which model has Thomas taken a photograph of him?
d. Which model have Thomas taken a photograph of him?
e. Which model Thomas has taken a photograph of?
f. Which model Thomas have taken a photograph of?
g. Which model Thomas has taken a photograph of him?
h. Which model Thomas have taken a photograph of him?

(B.26) **Lexicalizations**

a. Which model has Mary taken a photograph of?
b. Which opponent has Catherine destroyed a photograph of?
c. Which friend has Thomas painted a picture of?
d. Which enemy has John torn up a picture of?
e. Which celebrity has Sarah drawn a caricature of?
f. Which teacher has Betty ripped up a caricature of?
g. Which peasant has Peter done a painting of?
h. Which politician has Paul burnt a painting of?

The sentence in (B.27) was presented as the modulus. It contains violations of the constraints on verb class and definiteness for extraction from picture NPs.

(B.27) Which colleague has Terry torn up the picture of?

B.5. Experiment 5

The examples in (B.28) and (B.30) illustrate how the stimuli for Experiment 5 were constructed for the subexperiments on definiteness and verb class and on locality and referentiality. (B.29) and (B.31) list the lexicalizations that were used for the two subexperiments.

(B.28) **Definiteness and Verb Class**

a. HANNA saw a photograph of HER.
b. HANNA saw a photograph of HERSELF.
c. HANNA saw the photograph of HER.
d. HANNA saw the photograph of HERSELF.
e. EMILY took a photograph of HER.
f. EMILY took a photograph of HERSELF.
g. EMILY took the photograph of HER.
h. EMILY took the photograph of HERSELF.
i. RACHEL destroyed a photograph of HER.
j. RACHEL destroyed a photograph of HERSELF.
k. RACHEL destroyed the photograph of HER.

l. RACHEL destroyed the photograph of HERSELF.

(B.29) **Lexicalizations**

a. DAVID noticed a picture of HIM.
b. BILL painted a picture of HIM.
c. ROBERT tore up a picture of HIM.
d. REBECCA found a caricature of HER.
e. LIZ drew a caricature of HER.
f. ANNA ripped up a caricature of HER.
g. IAN discovered a painting of HIM.
h. BRIAN did a painting of HIM.
i. THOMAS burnt the painting of HIM.

(B.30) **Locality and Referentiality**

a. JULIA saw Peter's photograph of HER.
b. JULIA saw Peter's photograph of HERSELF.
c. Julia saw PETER'S photograph of HIM.
d. Julia saw PETER'S photograph of HIMSELF.
e. THE WOMAN saw Peter's photograph of HER.
f. THE WOMAN saw Peter's photograph of HERSELF.
g. The woman saw PETER'S photograph of HIM.
h. The woman saw PETER'S photograph of HIMSELF.
i. EACH WOMAN saw Peter's photograph of HER.
j. EACH WOMAN saw Peter's photograph of HERSELF.
k. Each woman saw PETER'S photograph of HIM.
l. Each woman saw PETER'S photograph of HIMSELF.

(B.31) **Lexicalizations**

a. ADAM noticed Sarah's picture of HIM.
b. THE MAN noticed Sarah's picture of HIM.
c. ALICE found Steven's caricature of HER.
d. THE GIRL found Steven's caricature of HER.
e. FRANK discovered Mary's painting of HIM.
f. THE BOY discovered Mary's painting of HIM.

The sentence in (B.32) was presented as the modulus. It contains the binding configuration anaphor-pronoun, and no c-command relationship holds between the anaphor and the pronoun.

(B.32) Jill told the people HE trusts all about SAM.

B.6. Experiment 6

Examples (B.33)–(B.36) give sample stimuli for the different pronominalizations used in Experiment 6. The lexicalizations used are listed in (B.37).

(B.33) **Non-Pronominalized Word Orders**

a. Ich weiß, dass der Kollege dem Handwerker den Kunden vermittelt.
I know that the colleague the craftsman the customer find
"I know that the colleague will find the customer for the craftsman."

b. Ich weiß, dass der Kollege den Kunden dem Handwerker vermittelt.

c. Ich weiß, dass dem Handwerker der Kollege den Kunden vermittelt.

d. Ich weiß, dass dem Handwerker den Kunden der Kollege vermittelt.

e. Ich weiß, dass den Kunden der Kollege dem Handwerker vermittelt.

f. Ich weiß, dass den Kunden dem Handwerker der Kollege vermittelt.

(B.34) **Word Orders with Subject Pronoun**

a. Ich weiß, dass er dem Handwerker den Kunden vermittelt.
I know that he the craftsman the customer find
"I know that he will find the customer for the craftsman."

b. Ich weiß, dass er den Kunden dem Handwerker vermittelt.

c. Ich weiß, dass dem Handwerker er den Kunden vermittelt.

d. Ich weiß, dass dem Handwerker den Kunden er vermittelt.

e. Ich weiß, dass den Kunden er dem Handwerker vermittelt.

f. Ich weiß, dass den Kunden dem Handwerker er vermittelt.

(B.35) **Word Orders with Indirect Object Pronoun**

a. Ich weiß, dass der Kollege ihm den Kunden vermittelt.
I know that the colleague him the customer find
"I know that the colleague will find the customer for him."

b. Ich weiß, dass der Kollege den Kunden ihm vermittelt.

c. Ich weiß, dass ihm der Kollege den Kunden vermittelt.

d. Ich weiß, dass ihm den Kunden der Kollege vermittelt.

e. Ich weiß, dass den Kunden der Kollege ihm vermittelt.

f. Ich weiß, dass den Kunden ihm der Kollege vermittelt.

(B.36) **Word Orders with Direct Object Pronoun**

a. Ich weiß, dass der Kollege dem Handwerker ihn vermittelt.
I know that the colleague the craftsman him find
"I know that he will find him for the craftsman."

b. Ich weiß, dass der Kollege ihn dem Handwerker vermittelt.

c. Ich weiß, dass dem Handwerker der Kollege ihn vermittelt.

d. Ich weiß, dass dem Handwerker ihn der Kollege vermittelt.

e. Ich weiß, dass ihn der Kollege dem Handwerker vermittelt.

f. Ich weiß, dass ihn dem Handwerker der Kollege vermittelt.

(B.37) **Lexicalizations**

a. Ich weiß, dass der Manager dem Projektleiter den Mitarbeiter vorstellt.
I know that the manager the project leader the employee introduces
"I know that the manager will introduce the employee to the project leader."

b. Ich glaube, dass der Produzent dem Regisseur den Schauspieler vorschlägt.
I know that the producer the director the actor proposes
"I know that the producer will propose the actor to the director."

c. Ich glaube, dass der Vater dem Onkel den Arzt empfiehlt.
I believe that the father the uncle the doctor recommends
"I believe that the father will recommend the doctor to the uncle."

d. Ich denke, dass der Soldat dem Offizier den Gefangenen übergibt.
I think that the soldier the officer the prisoner hands over
"I think that the soldier will hand over the prisoner to the officer."

e. Ich denke, dass der Zeuge dem Richter den Verdächtigen beschreibt.
I think that the witness the judge the suspect describes
"I think that the witness will describe the suspect to the judge."

f. Ich nehme an, dass der Täter dem Polizisten den Komplizen verrät.
I suppose that the offender the policeman the accomplice betrays
"I suppose that the offender will betray the accomplice to the policeman."

g. Ich nehme an, dass der Lehrer dem Direktor den Schüler schickt.
I suppose that the teacher the director the pupil sends
"I suppose that the teacher will send the pupil to the director."

This experiment used the same modulus as Experiment 2 (see (B.18)).

B.7. Experiment 7

The stimuli for the subexperiment are listed in (B.38), the corresponding contexts are in (B.39). The examples in (B.40) list sample stimuli for the subexperiment on ditransitive verbs; the lexicalizations are given in (B.41). The contexts for the ditransitive subexperiment is given in (B.42), with the lexicalizations in (B.43).

(B.38) **Transitive Verb Frames**

a. She repeated the question, and he the answer.

b. She negotiated with the manager, and he with the secretary.

c. She expected to win, and he to lose.

d. She read in the bedroom, and he in the lounge.

e. She shut the door, and he the window.

f. She traveled to Paris, and he to Vienna.

g. She agreed to stay, and he to leave.
h. She swam in the pool, and he in the sea.
i. She borrowed a book, and he a pencil.
j. She competed with a friend, and he with an enemy.
k. She intended to sleep, and he to stay awake.
l. She walked for two hours, and he for 30 minutes.
m. She bought a flat, and he a house.
n. She talked to the boss, and he to the employee.
o. She pretended to be happy, and he to be depressed.
p. She disappeared for a week, and he for two months.

(B.39) **Contexts**

a. What did Hanna and Michael repeat?
b. Who did Emily and Matthew negotiate with?
c. What did Rachel and Andrew expect to do?
d. Where did Rebecca and Mark read?
e. What did Liz and Joseph shut?
f. Where did Anna and Daniel travel to?
g. What did Julia and James agree to do?
h. Where did Alice and Charles swim?
i. What did Emma and David borrow?
j. Who did Laura and Bill compete with?
k. What did Monica and Robert intend to do?
l. How long did Maria and Ian walk?
m. What did Diana and Brian buy?
n. Who did Jenny and Steven talk to?
o. What did Jessica and Adam pretend to do?
p. How long did Miriam and Frank disappear for?

(B.40) **Ditransitive Verb Frames**

a. She charged the client 50 pounds, and he the manufacturer 100 pounds.
b. She charged the client 50 pounds, and the manufacturer 100 pounds.
c. She charged the client 50 pounds, and he 100 pounds.
d. She charged the client 50 pounds, and he the manufacturer.

(B.41) **Lexicalizations**

a. She accompanied the boy to school, and he the girl to university.
b. She authorized the manager to leave, and he the secretary to stay.
c. She wished the aunt merry Christmas, and he the uncle happy birthday.
d. She reported the spy to the CIA, and he the criminal to the FBI.

e. She preferred the relative to leave, and he the friend to stay.
f. She asked the friend a question, and he the colleague a favor.
g. She accused the girl of laziness, and he the boy of stupidity.
h. She forced the chairman to resign, and he the boss to retire.
i. She gave the sister a kiss, and he the brother a hug.
j. She threatened the president with a lawsuit, and he the boss with a strike.
k. She expected the mother to stay awake, and he the father to sleep.

(B.42) **Contexts**
a. Who did Hanna and Michael charge what?
b. Who did Hanna charge what?
c. what did Hanna and Michael charge the client?
d. Who did Hanna and Michael charge 50 pounds?

(B.43) **Lexicalizations**
a. Who did Emily and Matthew accompany where?
b. Who did Rachel and Andrew authorize to do what?
c. Who did Liz and Joseph wish what?
d. Who did Anna and Daniel report to who?
e. Who did Julia and James prefer to do what?
f. Who did Emma and David ask what?
g. Who did Laura and Bill accuse of what?
h. Who did Monica and Robert force to do what?
i. Who did Diana and Brian give what?
j. Who did Jenny and Steven threaten with what?
k. Who did Jessica and Adam expect to do what?

This experiment used the same modulus as Experiment 4 (see (B.27)).

B.8. Experiment 8

The examples in (B.44) list sample stimuli for Experiment 8; the lexicalizations are given in (B.45). The contexts used are given in (B.46), with the lexicalizations in (B.47).

(B.44) **Ditransitive Verb Frames**
a. He helped the neighbor by doing the shopping and the friend by washing the dishes.
b. He helped the neighbor by doing the shopping and she by washing the dishes.
c. He helped the neighbor by doing the shopping and the friend the dishes.
d. He helped the neighbor by doing the shopping and she the dishes.

e. He punished the criminal for robbing the bank and the thief for burgling the house.
f. He punished the criminal for robbing the bank and she for burgling the house.
g. He punished the criminal for robbing the bank and the thief the house.
h. He punished the criminal for robbing the bank and she the house.

(B.45) **Lexicalizations**

a. He discouraged the applicant by criticizing the plan and the candidate by rejecting the proposal.
b. He punished the criminal for robbing the bank and she the house.
c. He motivated the employee by praising the results and the colleague by approving the proposal.
d. He fined the driver for ignoring a one-way street and the cyclist for running a red light.
e. He annoyed the mother by wasting time and the father by squandering money.
f. He criticized the politician for committing an error and the minister for tolerating fraud.
g. He disrupted the neighbor by playing the piano and the flatmate by practicing the trumpet.
h. He praised the student for achieving a good result and the pupil for obtaining a high grade.
i. He upset the boy by breaking the bicycle and the girl by hiding the toy.
j. He scolded the boy for breaking a window and the girl for stealing a vase.
k. He treated the patient by giving an injection and the victim by prescribing drugs.
l. He thanked the neighbor for doing the shopping and the friend for washing the dishes.
m. He discredited the minister by uncovering fraud and the politician by exposing corruption.
n. He reprimanded the employee for ignoring a problem and the colleague for missing a deadline.

(B.46) **Contexts**

a. What happened?
b. Who did David help, and how?
c. How did David and Hanna help the neighbor?
d. How did David help the neighbor and the friend?
e. Who did David and Hanna help, and how?
f. Who did Michael punish, and why?
g. Why did Michael and Emma punish the criminal?
h. Why did Michael punish the criminal and the thief?
i. Who did Michael and Emma punish, and why?

(B.47) **Lexicalizations**

a. Who did Bill and Emily discourage, and how?
b. Who did Michael and Emma punish, and why?
c. Who did Robert and Rachel motivate, and how?
d. Who did Matthew and Laura fine, and why?
e. Who did Ian and Rebecca annoy, and how?
f. Who did Andrew and Monica criticism, and why?
g. Who did Brian disrupt, and how?
h. Who did Mark and Maria praise, and why?
i. Who did Steven and Anna upset, and how?
j. Who did Joseph and Diana scold, and why?
k. Who did Adam and Julia treat, and how?
l. Who did James and Jessica employ, and why?
m. Who did Frank and Alice discredit, and how?
n. Who did Charles and Miriam reprimand, and why?

The question-answer pair in (B.48) was presented as the modulus.

(B.48) Who did Peter and Mary blame? She blamed the manager for the mistake, and he the politician.

B.9. Experiment 9

The examples in (B.49) and (B.50) illustrate how the stimuli for Experiment 9 were constructed for the subexperiments on hard and soft constraint violations. (B.51) lists the lexicalizations, which were the same for both subexperiments.

(B.49) **Soft Constraint Violations**

a. Thomas seems to be very talented. Which model has he taken a photograph of?
b. Thomas has taken a photograph of one of his models. Which model has he taken a photograph of?
c. Thomas has taken this photograph of a model. Which model has he taken a photograph of?
d. Thomas takes a photograph of one of his models every week. Which model has he taken a photograph of this week?
e. Thomas seems to be very talented. How many models has he taken a photograph of?
f. Thomas seems to be very talented. Which model has he taken the photograph of?
g. Thomas seems to be very angry. Which model has he destroyed a photograph of?

h. Thomas has taken a photograph of some of his models. How many models has he taken a photograph of?
i. Thomas has taken this photograph of a model. Which model has he taken the photograph of?
j. Thomas destroys a photograph of one of his models every week. Which model has he destroyed a photograph of this week?

(B.50) **Hard Constraint Violations**

a. Thomas seems to be very talented. Which model he has taken a photograph of?
b. Thomas seems to be very talented. Which model has he taken a photograph of her?
c. Thomas seems to be very talented. Which model have he taken a photograph of?
d. Thomas has taken a photograph of one of his models. Which model he has taken a photograph of?
e. Thomas has taken a photograph of one of his models. Which model has he taken a photograph of her?
f. Thomas has taken a photograph of one of his models. Which model have he taken a photograph of?

(B.51) **Lexicalizations**

a. Mary seems to be very creative. Which friend has she painted a picture of?
b. Mary seems to be very aggressive. Which friend has she torn up a picture of?
c. Peter seems to be very imaginative. Which colleague has he drawn a caricature of?
d. Peter seems to be really furious. Which colleague has he ripped up a caricature of?
e. Sarah seems to be very creative. Which client has she done a painting of?
f. Sarah seems to be very annoyed. Which client has she burnt a painting of?

This experiment used the same modulus as Experiment 8 (see (B.48)).

B.10. Experiment 10

Examples (B.52)–(B.55) give sample stimuli for the different pronominalizations used in Experiment 10. The corresponding contexts are given in (B.56), the lexicalizations are listed in (B.57).

(B.52) **Non-Pronominalized Word Orders**

a. Maria glaubt, dass der Vater den Wagen kauft.
   Maria believes that the father the car buys
   “Maria believes that the father will buy the car.
b. Maria glaubt, dass den Wagen der Vater kauft.

c. Maria glaubt, dass kauft der Vater den Wagen.

d. Maria glaubt, dass kauft den Wagen der Vater.

(B.53) **Word Orders with Subject Pronoun**

a. Maria glaubt, dass er den Wagen kauft.
Maria believes that he the car buys
"Maria believes that he will buy the car.

b. Maria glaubt, dass den Wagen er kauft.

c. Maria glaubt, dass kauft er den Wagen.

d. Maria glaubt, dass kauft den Wagen er.

(B.54) **Word Orders with Object Pronoun**

a. Maria glaubt, dass der Vater ihn kauft.
Maria believes that the father it buys
"Maria believes that the father will buy it.

b. Maria glaubt, dass ihn der Vater kauft.

c. Maria glaubt, dass kauft der Vater ihn.

d. Maria glaubt, dass kauft ihn der Vater.

(B.55) **Word Orders with Subject and Object Pronoun**

a. Maria glaubt, dass er ihn kauft.
Maria believes that he it buys
"Maria believes that he will buy it.

b. Maria glaubt, dass ihn er kauft.

c. Maria glaubt, dass kauft er ihn.

d. Maria glaubt, dass kauft ihn er.

(B.56) **Contexts**

a. Was gibt's neues?
what is there new
"What's new?"

b. Wer kauft den Wagen?
who buys the car
"Who will buy the car?"

c. Was kauft der Vater?
what buys the father
"What will the father buy?"

(B.57) **Lexicalizations**

a. Petra weiß, dass der Junge den Brief schickt.
Petra knows that the boy the letter sends
"Petra knows that the boy will send the letter."

b. Klara sagt, dass der Chef den Bericht liest.
Klara says that the boss the report reads
“Klara know that the boss will read the report.”

c. Sabine behauptet, dass der Journalist den Artikel schreibt.
Sabine claims that the journalist the article writes
“Sabine claims that the journalist writes the article.”

d. Julia vermutet, dass der Soldat den Auftrag ablehnt.
Julia suspects that the soldier the order rejects
“Julia suspects that the soldier will reject the order.”

e. Tanja denkt, dass der Regisseur den Film dreht.
Tanja thinks that the director the film shots
“Tanja thinks that the director will shoot the film.”

f. Gabi nimmt an, dass der Anwalt den Vertrag formuliert.
Gabi supposes that the lawyer the contract formulates
“Gabi supposes that the lawyer will formulate the contract.”

g. Louise denkt, dass der Lehrer den Bus fährt.
Louise thinks that the teacher the bus drives
“Louise thinks that the teacher will drive the bus.”

This experiment used the same modulus as Experiment 1 (see (B.9)) for the null context condition, while for the context condition, the same modulus as in Experiment 3 (see (B.23)) was used.

B.11. Experiment 11

The sentences in (B.58) are sample stimuli for Experiment 11. The corresponding contexts are given in (B.59), the lexicalizations are listed in (B.60).

(B.58) **Word Orders**

a. O Tasos tha diavasi tin efimerida.
the Tasos will read the newspaper
“Tasos will read the newspaper.”

b. Tin efimerida tha diavasi o Tasos.

c. Tha diavasi o Tasos tin efimerida.

d. Tha diavasi tin efimerida o Tasos.

e. O Tasos tin efimerida tha diavasi.

f. Tin efimerida o Tasos tha diavasi.

(B.59) **Contexts**

a. Ti tha gini?
what will happen
“What will happen?

b. Pios tha diavasi tin efimerida?
who will read the newspaper
“Who will read the newspaper?”

c. Ti tha diavasi o Tasos?
what will read the Tasos
“What will Tasos read?”

d. Ti tha kani o Tasos me tin efimerida?
what will do the Tasos with the newspaper
“What will Tasos do with the newspaper?”

(B.60) **Lexicalizations**

a. O Petros tha plini to aftokinito.
the Petros will wash the car
“Petros will wash the car.”

b. O Kostas tha stili tin prosklisi.
the Kostas will send the invitation
“Kostas will send the invitation.”

c. O Tasos tha dhiavasi tin efimeridha.
the Tasos will read the newspaper
“Tasos will read the newspaper.”

d. O Giorgos tha grapsi to senario.
the Giorgos will write the script
“Giorgos will write the script.”

e. I Eleni tha pulisi tin afisa.
the Eleni will sell the poster
“Eleni will sell the poster.”

f. I Christina tha agorasi to pagoto.
the Christina will buy the ice-cream
“Christina will buy the ice-cream.”

g. I Maria tha dhi to spiti.
the Maria will see the house
“Maria will see the house.”

h. I Ana tha pari to leoforio.
the Ana will take the bus
“Ana will take the bus.”

The question-answer pair in (B.61) was presented as the modulus. The answer sentence is not fully grammatical, as the numeral *dodeka* should be preceded by the prepositional phrase *stis* instead of the definite article *tin.*

(B.61) Ti ora girise o Nikos htes? O Nikos girise tin dodeka.
which hour returned the Nikos yesterday the Nikos returned at twelve
“When did Nikos return yesterday? Nikos returned at twelve.”

B.12. Experiment 12

The sentences in (B.62) are sample stimuli for Experiment 12. The contexts and lexicalizations were the same as in Experiment 11, listed in (B.59) and (B.60).

(B.62) **Word Orders**

a. O Tasos tha tin diavasi tin efimerida.
the Tasos will it read the newspaper
"Tasos will read the newspaper."

b. Tin efimerida tha tin diavasi o Tasos.

c. Tha tin diavasi o Tasos tin efimerida.

The question-answer pair in (B.63) was presented as the modulus. The answer sentence is not fully grammatical, as the adverb *puthena* 'nowhere' should follow the verb, instead of preceding it.

(B.63) Pou tha tin pai tin Eleni o Vasilis? Den tha PUTHENA pai o Vasilis tin Eleni.
Where will her take the Eleni the Vasilis? Not will nowhere take the Vasilis the Eleni.
"Where will Vasilis take Eleni? Vasilis won't take Eleni anywhere."

Appendix C

Descriptive Statistics

The data were normalized by dividing each numeric judgment by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. Then the data were transformed by taking the decadic logarithm. This transformation ensures that the judgments are normally distributed and is standard practice for magnitude estimation data (Bard et al. 1996; Lodge 1981). All descriptive statistics were computed on the normalized, log-transformed judgments.

C.1. Experiment 1

Table C.1: Descriptives for Experiment 1, impersonal passive

dialect	verb class	mean	SD	SE
northern	change of location	.1869	.3268	.1033
	change of state	.0057	.2176	.0688
	continuation of state	−.0505	.4248	.1343
	existence of state (positional)	.2304	.2768	.0875
	controlled, non-motion	.3801	.3440	.1088
	controlled, motion	.2733	.3021	.0955
	uncontrolled, invol. reaction	.1440	.2098	.0663
	uncontrolled, emission	.2417	.1752	.0554
southern	change of location	.1417	.3875	.1225
	change of state	−.1580	.3333	.1054
	continuation of state	.2560	.4410	.1395
	existence of state (positional)	.1088	.2588	.0819
	controlled, non-motion	.3630	.3308	.1046
	controlled, motion	.3256	.2936	.0928
	uncontrolled, invol. reaction	.1748	.4182	.1323
	uncontrolled, emission	.1342	.3841	.1214

Table C.2: Descriptives for Experiment 1, auxiliary selection

dialect	verb class	auxiliary	mean	SD	SE
northern	change of location	haben	−.2288	.3545	.1121
		sein	.4148	.3237	.1024
	change of state	haben	−.2523	.3435	.1086
		sein	.3676	.4124	.1304
	continuation of state	haben	.3914	.3607	.1141
		sein	−.0949	.3823	.1209
	existence of state (positional)	haben	.2947	.3156	.0998
		sein	−.1890	.3422	.1082
	controlled, non-motion	haben	.4660	.3847	.1216
		sein	−.1965	.3382	.1070
	controlled, motion	haben	−.1036	.4247	.1343
		sein	.4366	.3837	.1213
	uncontrolled, invol. reaction	haben	.2420	.2875	.0909
		sein	.1181	.3675	.1162
	uncontrolled, emission	haben	.4027	.3601	.1139
		sein	−.2375	.3431	.1085
southern	change of location	haben	−.2100	.3798	.1201
		sein	.4300	.2428	.0768
	change of state	haben	−.1762	.3488	.1103
		sein	.4025	.3827	.1210
	continuation of state	haben	.2190	.2926	.0925
		sein	−.1905	.3496	.1105
	existence of state (positional)	haben	.1940	.3533	.1117
		sein	.1590	.1715	.0542
	controlled, non-motion	haben	.4520	.3856	.1219
		sein	−.1962	.3295	.1042
	controlled, motion	haben	.0321	.3343	.1057
		sein	.1162	.4082	.1291
	uncontrolled, invol. reaction	haben	.3451	.3430	.1085
		sein	−.1882	.3394	.1073
	uncontrolled, emission	haben	.3095	.2467	.0780
		sein	−.1141	.3490	.1104

C.2. Experiment 2

Table C.3: Descriptives for Experiment 2, impersonal passive

dialect	verb class	mean	SD	SE
northern	change of location, animate	.0333	.2490	.0719
	change of state, no prefix, inanimate	−.2841	.3483	.1006
	change of state, prefix, inanimate	−.3391	.3351	.0967
	continuation of state, inanimate	−.3684	.3494	.1009
	existence of state (positional), animate	.1339	.3034	.0876
	existence of state (positional), inanimate	−.2004	.3458	.0998
	controlled, non-motion, animate	.2312	.2859	.0825
	uncontr., invol. react., non-motion, animate	.0622	.5082	.1467
southern	change of location, animate	−.0668	.4064	.1127
	change of state, no prefix, inanimate	−.4617	.3689	.1023
	change of state, prefix, inanimate	−.5102	.4064	.1127
	continuation of state, inanimate	−.4672	.4171	.1157
	existence of state (positional), animate	−.0536	.4675	.1297
	existence of state (positional), inanimate	−.1052	.4716	.1308
	controlled, non-motion, animate	.1255	.3617	.1003
	uncontr., invol. react., non-motion, animate	−.1096	.3628	.1006

Table C.4: Descriptives for Experiment 2, auxiliary selection

dialect	verb class	auxiliary	mean	SD	SE
northern	change of location, animate	haben	−.4451	.3635	.1049
		sein	.3292	.2909	.0840
	change of state, no prefix, inanimate	haben	−.0946	.5627	.1624
		sein	.2590	.3675	.1061
	change of state, prefix, inanimate	haben	−.3587	.3783	.1092
		sein	.2546	.2792	.0806
	continuation of state, inanimate	haben	.1797	.3132	.0904
		sein	−.3841	.4070	.1175
	existence of state (positional), animate	haben	.1421	.4034	.1165
		sein	−.1418	.5224	.1508
	existence of state (positional), inanimate	haben	.1946	.3642	.1051
		sein	−.0786	.4768	.1377
	controlled, non-motion, animate	haben	.3064	.2894	.0836
		sein	−.5284	.2327	.0672
	uncontr., invol. react., non-motion, animate	haben	.2627	.3444	.0994
		sein	−.4210	.5275	.1523
southern	change of location, animate	haben	−.3149	.4071	.1129
		sein	.2818	.2157	.0598
	change of state, no prefix, inanimate	haben	.0382	.3717	.1031
		sein	−.0954	.7114	.1973
	change of state, prefix, inanimate	haben	−.5289	.5196	.1441
		sein	.2465	.1581	.0438
	continuation of state, inanimate	haben	.2563	.2506	.0695
		sein	−.4678	.3105	.0861
	existence of state (positional), animate	haben	.1662	.1858	.0515
		sein	.1697	.3494	.0969
	existence of state (positional), inanimate	haben	.1448	.1395	.0387
		sein	.0808	.1913	.0531
	controlled, non-motion, animate	haben	.2771	.2201	.0610
		sein	−.4821	.5607	.1555
	uncontr., invol. react., non-motion, animate	haben	.2065	.3032	.0841
		sein	−.5141	.5442	.1509

C.3. Experiment 3

Table C.5: Descriptives for Experiment 3

dialect	verb class	telicity	auxiliary	mean	SD	SE
northern	continuation of state	telic	haben	.3001	.2226	.0671
			sein	−.0822	.2169	.0654
		atelic	haben	.3090	.2439	.0735
			sein	−.1684	.2966	.0894
	existence of state (positional)	telic	haben	.1406	.3759	.1133
			sein	.0207	.2188	.0660
		atelic	haben	.2820	.3385	.1021
			sein	.0161	.2241	.0676
	controlled, motion	telic	haben	−.3580	.3156	.0951
			sein	.2836	.2851	.0859
		atelic	haben	−.0019	.2625	.0792
			sein	.0671	.3295	.0993
	uncontrolled, emission	telic	haben	−.1219	.2662	.0803
			sein	.2495	.2634	.0794
		atelic	haben	.1906	.1084	.0327
			sein	.0167	.1720	.0519
southern	continuation of state	telic	haben	.3399	.2913	.0808
			sein	−.0228	.4035	.1119
		atelic	haben	.3038	.3693	.1024
			sein	−.0832	.2740	.0760
	existence of state (positional)	telic	haben	.2060	.4096	.1136
			sein	.2683	.4134	.1147
		atelic	haben	.1865	.3985	.1105
			sein	.1875	.3991	.1107
	controlled, motion	telic	haben	−.1385	.3619	.1004
			sein	.4159	.3702	.1027
		atelic	haben	.1639	.3140	.0871
			sein	.2614	.3966	.1100
	uncontrolled, emission	telic	haben	−.2076	.4211	.1168
			sein	.2394	.4953	.1374
		atelic	haben	.1394	.3994	.1108
			sein	−.1893	.4874	.1352

C.4. Experiment 4

Table C.6: Descriptives for Experiment 4, soft constraint violations

verb class	referentiality	definiteness	mean	SD	SE
[−EXISTENCE]	which	indefinite	.0865	.1660	.0326
		definite	.0473	.1344	.0264
	how many	indefinite	.0877	.1598	.0313
		definite	.0016	.1237	.0243
[+EXISTENCE]	which	indefinite	.0540	.1295	.0254
		definite	.0032	.1097	.0215
	how many	indefinite	−.0487	.2218	.0435
		definite	−.0727	.1674	.0328

Table C.7: Descriptives for Experiment 4, hard constraint violations

inversion	resumptive	agreement	mean	SD	SE
yes	no	yes	.0382	.1878	.0368
		no	−.2527	.2038	.0400
	yes	yes	−.3746	.2788	.0547
		no	−.4170	.2822	.0553
no	no	yes	−.2217	.3363	.0660
		no	−.3604	.2665	.0523
	yes	yes	−.3696	.2525	.0495
		no	−.4486	.2997	.0588

C.5. Experiment 5

Table C.8: Descriptives for Experiment 5, definiteness and verb class

verb class	definiteness	anaphor	mean	SD	SE
achievement	indefinite	pronoun	.2316	.5479	.0760
[−EXISTENCE]		reflexive	.6755	.5311	.0737
	definite	pronoun	.3058	.5578	.0774
		reflexive	.6483	.5532	.0767
accomplishment	indefinite	pronoun	.0063	.4341	.0602
[−EXISTENCE]		reflexive	.6801	.5290	.0734
	definite	pronoun	.1318	.5083	.0705
		reflexive	.6677	.5435	.0754
accomplishment	indefinite	pronoun	.1989	.5327	.0739
[+EXISTENCE]		reflexive	.6737	.5394	.0748
	definite	pronoun	.2982	.5322	.0738
		reflexive	.6757	.5257	.0729

Table C.9: Descriptives for Experiment 5, locality and referentiality

binder	referentiality	anaphor	mean	SD	SE
remote	proper name	pronoun	.4999	.5287	.0733
		reflexive	.5407	.5939	.0824
	definite NP	pronoun	.4907	.5153	.0715
		reflexive	.5520	.5126	.0711
	quantified NP	pronoun	.3203	.5389	.0747
		reflexive	.4858	.5348	.0742
local	proper name	pronoun	.1754	.5379	.0746
		reflexive	.6530	.5235	.0726
	definite NP	pronoun	.1798	.5933	.0823
		reflexive	.6457	.5253	.0729
	quantified NP	pronoun	.1944	.6546	.0908
		reflexive	.6027	.5618	.0779

C.6. Experiment 6

Table C.10: Descriptives for Experiment 6

word order	mean	SD	SE
SIOV	.2083	.2695	.0539
SOIV	.0994	.3259	.0652
ISOV	−.0716	.1918	.0384
IOSV	−.2667	.2755	.0551
OSIV	−.2038	.3079	.0616
OISV	−.2736	.2750	.0550
S_{pro}IOV	.1386	.3445	.0689
S_{pro}OIV	.1519	.2493	.0499
IS_{pro}OV	−.1463	.2545	.0509
IOS_{pro}V	−.2936	.2426	.0485
OS_{pro}IV	−.2081	.2277	.0455
OIS_{pro}V	−.3471	.2676	.0535
SI$_{pro}$OV	.1471	.3138	.0628
SOI$_{pro}$V	−.0516	.2991	.0598
I$_{pro}$SOV	.1144	.2491	.0498
I$_{pro}$OSV	−.2612	.3113	.0623
OSI$_{pro}$V	−.2810	.3102	.0620
OI$_{pro}$SV	−.2164	.2458	.0492
SIO$_{pro}$V	−.1876	.2607	.0521
SO$_{pro}$IV	.1938	.2651	.0530
ISO$_{pro}$V	−.2694	.2351	.0470
IO$_{pro}$SV	−.3550	.3406	.0681
O$_{pro}$SIV	.1235	.3101	.0620
O$_{pro}$ISV	−.2247	.2591	.0518

C.7. Experiment 7

Table C.11: Descriptives for Experiment 7

context	verb frame	mean	SD	SE
null	NP V NP	.1182	.1793	.0352
	NP V PP	.1144	.2383	.0467
	NP V VP	.1130	.1892	.0371
	NP V PP-adj	.1571	.1907	.0374
felicitous	NP V NP	.0969	.1404	.0293
	NP V PP	.0873	.1273	.0265
	NP V VP	.1236	.1078	.0225
	NP V PP-adj	.0909	.1451	.0303

C.8. Experiment 8

Table C.12: Descriptives for Experiment 8

context	verb class	verb frame	mean	SD	SE
null	[+CONTROL]	_ _ NP [V NP]	.0682	.1777	.0355
		_ _ NP [_ NP]	-.1757	.3198	.0640
	[−CONTROL]	NP _ _ [V NP]	-.1226	.2697	.0539
		NP _ _ [_ NP]	-.2140	.2630	.0526
		_ _ NP [V NP]	.0099	.2070	.0414
		_ _ NP [_ NP]	-.1782	.2741	.0548
	[+CONTROL]	NP _ _ [V NP]	-.1246	.2411	.0482
		NP _ _ [_ NP]	-.2808	.3029	.0606
neutral	[+CONTROL]	_ _ NP [V NP]	.1058	.1377	.0251
		_ _ NP [_ NP]	−.1137	.2211	.0404
	[−CONTROL]	NP _ _ [V NP]	−.0609	.2472	.0451
		NP _ _ [_ NP]	−.2191	.2429	.0443
		_ _ NP [V NP]	.0096	.2086	.0381
		_ _ NP [_ NP]	−.1572	.2733	.0499
	[+CONTROL]	NP _ _ [V NP]	−.1422	.2136	.0390
		NP _ _ [_ NP]	−.1684	.2434	.0444
felicitous	[+CONTROL]	_ _ NP [V NP]	.1274	.1915	.0350
		_ _ NP [_ NP]	−.1040	.2111	.0385
	[−CONTROL]	NP _ _ [V NP]	.0869	.2446	.0447
		NP _ _ [_ NP]	−.0924	.1902	.0347
		_ _ NP [V NP]	.0837	.2049	.0374
		_ _ NP [_ NP]	−.1554	.2281	.0416
	[+CONTROL]	NP _ _ [V NP]	−.0278	.1700	.0310
		NP _ _ [_ NP]	−.1082	.1883	.0344
non-felicitous	[+CONTROL]	_ _ NP [V NP]	.1497	.1494	.0273
		_ _ NP [_ NP]	−.1016	.2639	.0482
	[−CONTROL]	NP _ _ [V NP]	−.0081	.2114	.0386
		NP _ _ [_ NP]	−.1299	.1776	.0324
		_ _ NP [V NP]	.0986	.2081	.0380
		_ _ NP [_ NP]	−.1185	.2396	.0437
	[+CONTROL]	NP _ _ [V NP]	−.0941	.2153	.0393
		NP _ _ [_ NP]	−.1434	.1853	.0338

C.9. Experiment 9

Table C.13: Descriptives for Experiment 9, soft constraint violations

context	verb class	referentiality	definiteness	mean	SD	SE
neutral	[−EXISTENCE]	which	indefinite	.0323	.2997	.0547
	[+EXISTENCE]	which	indefinite	−.0289	.2951	.0539
	[−EXISTENCE]	which	indefinite	.0323	.2997	.0547
	[−EXISTENCE]	how many	indefinite	.0154	.2875	.0525
	[−EXISTENCE]	which	indefinite	.0323	.2997	.0547
	[−EXISTENCE]	which	definite	.0064	.3114	.0569
felicitous	[−EXISTENCE]	which	indefinite	.1119	.2736	.0500
	[+EXISTENCE]	which	indefinite	.0476	.2758	.0503
	[−EXISTENCE]	which	indefinite	.0491	.2480	.0453
	[−EXISTENCE]	how many	indefinite	.0544	.3134	.0572
	[−EXISTENCE]	which	indefinite	.0424	.2673	.0488
	[−EXISTENCE]	which	definite	.0345	.2895	.0528

Table C.14: Descriptives for Experiment 9, hard constraint violations

context	inversion	resumptive	agreement	mean	SD	SE
neutral	yes	no	yes	.0323	.2997	.0547
	no	no	yes	−.0756	.2756	.0503
	yes	no	yes	.0323	.2997	.0547
	yes	yes	yes	−.3442	.2065	.0377
	yes	no	yes	.0323	.2997	.0547
	yes	no	no	−.3100	.3389	.0619
felicitous	yes	no	yes	.0491	.2480	.0453
	no	no	yes	−.0365	.2843	.0519
	yes	no	yes	.0491	.2480	.0453
	yes	yes	yes	−.3697	.2832	.0517
	yes	no	yes	.0491	.2480	.0453
	yes	no	no	−.2244	.3318	.0606

C.10. Experiment 10

Table C.15: Descriptives for Experiment 10, null context condition

word order	mean	SD	SE
SOV	.3818	.3232	.0723
OSV	.1078	.3154	.0705
VSO	−.2228	.2772	.0620
VOS	−.1900	.2287	.0511
$S_{pro}OV$	.4180	.3266	.0730
$OS_{pro}V$	−.0887	.2224	.0497
$VS_{pro}O$	−.1861	.2255	.0504
VOS_{pro}	−.2188	.2667	.0596
$SO_{pro}V$	.2482	.2960	.0662
$O_{pro}SV$	.2412	.3456	.0773
VSO_{pro}	−.2153	.2435	.0544
$VO_{pro}S$	−.2115	.2435	.0544
$S_{pro}O_{pro}V$	.3024	.3211	.0718
$O_{pro}S_{pro}V$	−.1071	.3406	.0762
$VS_{pro}O_{pro}$	−.1992	.2576	.0576
$VO_{pro}S_{pro}$	−.2599	.2619	.0586

Table C.16: Descriptives for Experiment 10, context condition

context	word order	mean	SD	SE
all focus	SOV	.3848	.3042	.0546
	OSV	.1110	.2612	.0469
	VSO	−.1871	.2532	.0455
	VOS	−.2396	.2455	.0441
	$SO_{pro}V$	.1002	.3234	.0581
	$O_{pro}SV$	.0215	.3272	.0588
	VSO_{pro}	−.3903	.3488	.0627
	$VO_{pro}S$	−.2590	.3084	.0554
S focus	SOV	.4668	.2245	.0403
	OSV	.2112	.2034	.0365
	VSO	−.1247	.2698	.0485
	VOS	−.1898	.2244	.0403
	$SO_{pro}V$	.4252	.2281	.0410
	$O_{pro}SV$	.3957	.2671	.0480
	VSO_{pro}	−.1654	.3116	.0560
	$VO_{pro}S$	−.1273	.3046	.0547
O focus	SOV	.3752	.2569	.0461
	OSV	.1346	.2385	.0428
	VSO	−.1918	.2622	.0471
	VOS	−.1696	.2208	.0397
	$S_{pro}OV$	.3808	.1890	.0339
	$OS_{pro}V$	.0157	.2631	.0473
	$VS_{pro}O$	−.1322	.2678	.0481
	VOS_{pro}	−.1909	.2905	.0522

C.11. Experiment 11

Table C.17: Descriptives for Experiment 11, null context condition

word order	mean	SD	SE
SVO	.5660	.3588	.0870
OVS	.4285	.3434	.0833
VSO	.4830	.3614	.0877
VOS	.4679	.3487	.0846
SOV	.3637	.3112	.0755
OSV	.3455	.3387	.0822

Table C.18: Descriptives for Experiment 11, context condition

context	word order	mean	SD	SE
all focus	SVO	.3817	.3312	.0803
	OVS	.0673	.3556	.0863
	VSO	.2658	.2401	.0582
	VOS	.2362	.2680	.0650
	SOV	.1185	.2969	.0720
	OSV	.0651	.2925	.0709
S focus	SVO	.3619	.2802	.0680
	OVS	.2526	.2637	.0640
	VSO	.2568	.2978	.0722
	VOS	.1933	.3304	.0801
	SOV	.0738	.2149	.0521
	OSV	.1225	.2490	.0604
	SVO	.3626	.2870	.0696
O focus	OVS	.2299	.2694	.0653
	VSO	.1484	.2652	.0643
	VOS	.2407	.2087	.0506
	SOV	.1583	.3024	.0733
	OSV	.1320	.2769	.0672
	SVO	.3168	.3194	.0775
V focus	OVS	.0877	.3400	.0825
	VSO	.1311	.2681	.0650
	VOS	.2181	.2780	.0674
	SOV	.1160	.2299	.0558
	OSV	.0383	.1972	.0478

C.12. Experiment 12

Table C.19: Descriptives for Experiment 12, null context condition

word order	mean	SD	SE
Svo	.3475	.1991	.0575
svO	.4373	.1708	.0493
ovS	.2604	.2204	.0636
Ovs	.2777	.1751	.0505
vSo	.3096	.2224	.0642
vsO	.3441	.1909	.0551
Sclvo	.3524	.1888	.0545
sclvO	.1943	.2006	.0579
oclvS	.3707	.1875	.0541
Oclvs	.1728	.2097	.0605
clvSo	.3453	.2230	.0644
clvsO	.2323	.2004	.0579

Table C.20: Descriptives for Experiment 12, context condition

context	word order	mean	SD	SE
all focus	Svo	.3506	.2732	.0663
	svO	.4112	.2278	.0552
	ovS	.1892	.2900	.0703
	Ovs	.1728	.3941	.0956
	vSo	.3318	.3547	.0860
	vsO	.4254	.2053	.0498
	Sclvo	.2754	.4216	.1023
	sclvO	.2896	.2565	.0622
	oclvS	.3912	.2006	.0487
	Oclvs	.1119	.3754	.0910
	clvSo	.3720	.2577	.0625
	clvsO	.2802	.2572	.0624
S focus	Svo	.5261	.2412	.0585
	svO	.3790	.2408	.0584
	ovS	.4077	.1811	.0439
	Ovs	.1474	.2251	.0546
	vSo	.4096	.2083	.0505
	vsO	.2756	.2419	.0587
	Sclvo	.4972	.2359	.0572
	sclvO	.2685	.2280	.0553
	oclvS	.4892	.2651	.0643
	Oclvs	.1758	.2659	.0645
	clvSo	.5155	.2105	.0510
	clvsO	.2695	.2330	.0565
O focus	Svo	.2015	.3553	.0862
	svO	.5391	.2091	.0507
	ovS	.2716	.2733	.0663
	Ovs	.4665	.3926	.0952
	vSo	.1181	.4128	.1001
	vsO	.3307	.2316	.0562
	Sclvo	.0797	.3289	.0798
	sclvO	.1156	.3251	.0788
	oclvS	.2285	.2669	.0647
	Oclvs	.2513	.3029	.0735
	clvSo	−.0165	.3052	.0740
	clvsO	.1180	.2705	.0656
V focus	Svo	.2682	.2511	.0609
	svO	.4039	.2962	.0718
	ovS	.1369	.3374	.0818
	Ovs	.1813	.2559	.0621
	vSo	.2083	.2576	.0625
	vsO	.3191	.2731	.0662
	Sclvo	.1789	.3775	.0916
	sclvO	.2836	.2358	.0572
	oclvS	.1811	.4277	.1037
	Oclvs	.0833	.3596	.0872
	clvSo	.2782	.2607	.0632
	clvsO	.1618	.4067	.0986

C.13. Experiment 13

Table C.21: Descriptives for Experiment 13

verb class	referentiality	definiteness	mean	SD	SE
[−EXISTENCE]	which	indefinite	.0927	.2226	.0445
		definite	.0754	.2170	.0434
	how many	indefinite	.0360	.3050	.0610
		definite	−.0398	.2674	.0535
[+EXISTENCE]	which	indefinite	.0185	.2999	.0600
		definite	−.0101	.2938	.0588
	how many	indefinite	−.0603	.2541	.0508
		definite	−.0726	.2738	.0548

C.14. Experiment 14

Table C.22: Descriptives for Experiment 14, first subexperiment

antecedent	NP_1	NP_2	c-command	mean	SD	SE
subject	name	pronoun	no	.6321	.5493	.1373
			yes	.5124	.6274	.1569
	name	name	no	.3566	.5406	.1351
			yes	.6786	.5311	.1328
	pronoun	name	no	.3309	.4853	.1213
			yes	.0456	.4925	.1231
object	name	pronoun	no	.5230	.6009	.1502
			yes	.4143	.5050	.1262
	name	name	no	.2686	.5296	.1324
			yes	.6569	.5557	.1389
	pronoun	name	no	.4584	.5791	.1448
			yes	.0891	.3907	.0977

Table C.23: Descriptives for Experiment 14, second subexperiment

NP_1	NP_2	c-command	mean	SD	SE
name	pronoun	no	.5451	.2962	.0541
		yes	.4677	.3825	.0698
name	name	no	.5940	.3198	.0584
		yes	.6136	.3084	.0563
pronoun	name	no	.3532	.4812	.0879
		yes	.0847	.3908	.0714

Table C.24: Descriptives for Experiment 14, third subexperiment

NP_1	NP_2	c-command	mean	SD	SE
name	pronoun	no	.3801	.5224	.1349
pronoun	name	no	.1140	.3568	.0921
name	name	no	.5468	.6546	.1690
pronoun	anaphor	no	−.1302	.4577	.1182
name	anaphor	no	−.0930	.4323	.1116
name	pronoun	yes	−.1794	.4662	.1204
name	name	yes	−.2676	.3480	.0899
pronoun	name	yes	.6185	.6441	.1663

Table C.25: Descriptives for Experiment 14, fourth subexperiment

antecedent	NP_1	NP_2	c-command	mean	SD	SE
subject	name	pronoun	no	.2489	.4288	.1107
			yes	.4977	.5244	.1354
	name	name	no	.1793	.5987	.1546
			yes	.5185	.5427	.1401
	pronoun	name	no	.3940	.4854	.1253
			yes	.0542	.4667	.1205
object	name	pronoun	no	.3772	.4661	.1203
			yes	.4981	.5608	.1448
	name	name	no	.2557	.6796	.1755
			yes	.5550	.5306	.1370
	pronoun	name	no	.4611	.5039	.1301
			yes	−.0187	.3880	.1002

Bibliography

Abney, Steven. 1996. Statistical methods and linguistics. In Judith Klavans and Philip Resnik, eds., *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA: MIT Press.

—. 1997. Stochastic attribute-value grammars. *Computational Linguistics* 23(4): 597–618.

Agouraki, Yeoryia. 1993. *Spec-Head Licensing: The Scope of the Theory*. Ph.D. thesis, University College London.

Alexopoulou, Theodora. 1998. *The Syntax of Discourse Functions in Greek: A Non-configurational Approach*. Ph.D. thesis, University of Edinburgh.

Altmann, Gerry T. M., and Mark J. Steedman. 1988. Interaction with context during human sentence processing. *Cognition* 30(3): 191–238.

Anagnostopoulou, Elena. 1994. *Clitic Dependencies in Modern Greek*. Ph.D. thesis, University of Salzburg.

Anttila, Arto. 1997. Deriving variation from grammar: A study of Finnish genitives. In Frans Hinskens, Roeland van Hout, and W. Leo Wetzels, eds., *Variation, Change, and Phonological Theory*. Amsterdam: John Benjamins.

Aoun, Joseph, Norbert Hornstein, David Lightfoot, and Amy Weinberg. 1987. Two types of locality. *Linguistic Inquiry* 18(4): 537–577.

Asudeh, Ash. 1998. *Anaphora and Argument Structure: Topics in the Syntax and Semantics of Reflexives and Reciprocals*. Master's thesis, University of Edinburgh.

—. 2001. Linking, optionality, and ambiguity in Marathi. In Peter Sells, ed., *Formal and Empirical Issues in Optimality-theoretic Syntax*. Stanford, CA: CSLI Publications.

Bader, Markus, and Michael Meng. 1999. Subject-object ambiguities in German embedded clauses: An across-the-board comparison. *Journal of Psycholinguistic Research* 28(2): 121–143.

Barbosa, Pilar, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky, eds. 1998. *Is the Best Good Enough? Optimality and Competition in Syntax*. Cambridge, MA: MIT

Press and MIT Working Papers in Linguistics.

Bard, Ellen Gurman, Cheryl Frenck-Mestre, Louise Kelly, Kerry Killborn, and Antonella Sorace. 1999. Judgement and perception of gradable linguistic anomaly. Unpubl. ms., Human Communication Research Centre, University of Edinburgh.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1): 32–68.

Belletti, Adriana, Luciana Brandi, and Luigi Rizzi, eds. 1981. *Theory of Markedness in Generative Grammar: Proceedings of the 4th GLOW Conference*. Scuole Normale Superiore di Pisa.

Belletti, Adriana, and Luigi Rizzi. 1988. Psych-verbs and θ-theory. *Natural Language and Linguistic Theory* 6(3): 291–352.

Birch, Stacy, and Charles Clifton, Jr. 1995. Focus, accent, and argument structure: Effects on language comprehension. *Language and Speech* 38(4): 365–391.

Birdsong, David. 1989. *Metalinguistic Performance and Interlinguistic Competence*. Berlin: Springer.

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences*, vol. 21, 43–58. University of Amsterdam.

—. 1998. *Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives*. The Hague: Holland Academic Graphics.

—. 1999a. Optimality-theoretic learning in the Praat program. In *Proceedings of the Institute of Phonetic Sciences*, vol. 23, 17–35. University of Amsterdam.

—. 1999b. Review of: Arto Anttila (1997): Variation in Finnish phonology and morphology. Unpubl. ms., Institute of Phonetic Sciences, University of Amsterdam.

—. 2000. Learning a grammar in functional phonology. In Dekkers, van der Leeuw, and van de Weijer 2000, 465–523.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1).

Boersma, Paul, and Clara Levelt. 1999. Gradual constraint-ranking learning algorithm predicts acquisition order. In Eve V. Clark, ed., *Proceedings of the 30th Child Language Research Forum*. Stanford, CA: CSLI Publications.

Bolinger, Dwight L. 1961a. *Generality, Gradience, and the All-Or-None*. The Hague: Mouton.

—. 1961b. Syntactic blends and other matters. *Language* 37: 366–381.

—. 1978. Asking more than one thing at a time. In H. Hiz, ed., *Questions*, 87–106. Dordrecht: Reidel.

—. 1989. *Intonation and its uses: melody in grammar and discourse*. Stanford: Stanford University Press.

Bresnan, Joan. 2000. Optimal syntax. In Dekkers et al. 2000, 334–385.

Brew, Chris. 1994. Comments on Eisele: Types and clauses: Two styles of probabilistic processing in CUF. In Dörre 1994, 23–28.

—. 1995. Stochastic HPSG. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 83–89. Dublin.

Carlson, Katy. 1999. The effects of parallelism and prosody in the processing of gapping structures. Unpubl. ms., Department of Linguistics, University of Massachusetts, Amherst.

Carroll, John, ed. 1996. *Proceedings of the Workshop on Robust Parsing*, 8th European Summer School in Logic, Language and Information, Prague.

Cedergren, Henrietta J., and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50(2): 333–355.

Chafe, W. L. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Charles N. Li, ed., *Subject and Topic*, 25–55. New York: Academic Press.

—. 1983. *Meaning and the Structure of Language.* Chicago: University of Chicago Press.

Chapman, Robin S. 1974. *The Interpretation of Deviant Sentences in English: A Transformational Approach.* The Hague: Mouton.

Chaudron, C. 1983. Research on metalinguistic judgments: A review of theory, methods, and results. *Language Learning* 33(3): 343–377.

Choi, Hye-Won. 1996. *Optimizing Structure in Context: Scrambling and Information Structure*. Ph.D. thesis, Stanford University.

Chomsky, Noam. 1955. The logical structure of linguistic theory. Ms., Harvard University and MIT. Published as Chomsky 1975.

—. 1964. Degrees of grammaticalness. In Fodor and Katz 1964, 384–389.

—. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

—. 1975. *The Logical Structure of Linguistic Theory*. New York: Plenum Press.

—. 1981. *Lectures on Government and Binding.* Dordrecht: Foris.

—. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

—. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam, and Howard Lasnik. 1995. The theory of principles and parameters. In Chomsky 1995, 13–127.

COGSCI. 1990. *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Lawrence Erlbaum Associates.

Coleman, E. B. 1965. Responses to a scale of grammaticalness. *Journal of Verbal Learning and Verbal Behavior* 4: 521–527.

Core, Mark G. 1999. *Dialog Parsing: From Speech Repairs to Speech Acts*. Ph.D. thesis, University of Rochester.

Cowart, Wayne. 1989a. Illicit acceptability in *picture* NPs. In Caroline Wiltshire, Randolph Graczyk, and Bradley Music, eds., *Papers from the 25th Meeting of the Chicago Linguistic Society*, vol. 1: The General Session, 27–40. Chicago.

—. 1989b. Notes on the biology of syntactic processing. *Journal of Psycholinguistic Research* 18(1): 89–103.

—. 1994. Anchoring and grammar effects in judgments of sentence acceptability. *Perceptual and Motor Skills* 79(3): 1171–1182.

—. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

Cowart, Wayne, G. Andrew Smith-Petersen, and Sadie Fowler. 1998. Equivocal evidence on field-dependence effects in sentence judgments. *Perceptual and Motor Skills* 87: 1091–1102.

Culy, Christopher. 1998. Statistical distribution and the grammatical/ungrammatical distinction. *Grammars* 1(1): 1–19.

Dekkers, Joost, Frank van der Leeuw, and Jeroen van de Weijer, eds. 2000. *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford: Oxford Univeristy Press.

Diesing, Molly. 1992. *Indefinites*. Cambridge, MA: MIT Press.

Dörre, Jochen, ed. 1994. *Computational Aspects of Constraint-Based Linguistic Description*. No. R1.2.B in DYANA-2 Deliverables. ILLC/Department of Philosophy, University of Amsterdam.

Edwards, Allen L. 1984. *An Introduction to Linear Regression and Correlation*. New York: W. H. Freeman, 2nd edn.

Eisele, Andreas. 1994. Towards probabilistic extensions of constraint-based grammars. In Dörre 1994, 1–22.

Erbach, Gregor. 1993. Towards a theory of degrees of grammaticality. CLAUS Report 34, Department of Computational Linguistics, Saarland University.

—. 1997. *Bottom-Up Earley Deduction for Preference-Driven Natural Language Processing*, vol. 4 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI, Saarland University.

Erteschik-Shir, Nomi. 1981. On extraction from noun phrases (picture noun phrases). In Belletti, Brandi, and Rizzi 1981, 147–169.

Fiengo, Robert. 1987. Definiteness, specificity, and familiarity. *Linguistic Inquiry* 18: 163–166.

Fodor, Jerry A., and Jerrold J. Katz, eds. 1964. *The Structure of Language: Readings in the Philosophy of Language*. Englewood Cliffs, NJ: Prentice-Hall.

Gibson, E., and J. Thomas. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes* 14(3): 225–248.

Gordon, Peter C., and Randall Hendrick. 1997. Intuitive knowledge of linguistic co-reference. *Cognition* 62: 325–370.

—. 1998a. Dimensions of grammatical coreference. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, 424–429. Mahwah, NJ: Lawrence Erlbaum Associates.

—. 1998b. Nondefinite NP anaphora: A reappraisal. In M. Catherine Gruber, Derrick Higgins, Kenneth S. Olson, and Tamra Wysocki, eds., *Papers from the 34th Meeting of the Chicago Linguistic Society*, vol. 1: The Main Session, 195–210. Chicago.

—. 1998c. Representation and processing of coreference in discourse. *Cognitive Science* 22(4): 389–424.

Greenbaum, Sidney. 1973. Informant elicitation of data on syntactic variation. *Lingua* 31: 201–212.

—. 1976. Contextual influence on acceptability judgments. *Linguistics* 187: 5–11.

—. 1977. Judgments of syntactic acceptability and frequency. *Studia Linguistica* 31(2): 83–105.

Greenbaum, Sidney, and Charles F. Meyer. 1982. Ellipsis and coordination: Norms and preferences. *Language and Communication* 2: 231–240.

Greenbaum, Sidney, and Randolph Quirk. 1970. *Elicitation Experiments in English: Linguistic Studies in Use and Attitude*. London: Longman.

Grewendorf, Günther. 1989. *Ergativity in German*. Dordrecht: Foris.

Grimshaw, Jane. 1997. Projection, heads, and optimality. *Linguistic Inquiry* 28: 373–422.

Gruber, M. Catherine, Derrick Higgins, Kenneth S. Olson, and Tamra Wysocki, eds. 1998. *Papers from the 34th Meeting of the Chicago Linguistic Society*, vol. 2: The Panels, Chicago.

Guy, Gregory R. 1997. Violable is variable: Optimality Theory and linguistic variation. *Language Variation and Change* 9: 333–347.

Guy, Gregory R., and Charles Boberg. 1997. Inherent variability and the obligatory contour principle. *Language Variation and Change* 9: 149–164.

Haegeman, Liliane. 1994. *Introduction to Government and Binding Theory*. Oxford: Basil Blackwell, 2nd edn.

Haider, Hubert. 1993. *Deutsche Syntax generativ: Vorstudien zur Theorie einer projektiven Grammatik*. Tübingen: Gunter Narr.

Haider, Hubert, and Rositta Rindler-Schjerve. 1987. The parameter of auxiliary selection: Italian-German contrasts. *Linguistics* 25: 1029–1055.

Halliday, Michael A. K. 1967. Notes on transitivity and theme in English, part II. *Journal of Linguistics* 3: 199–244.

Hankamer, Jorge. 1973. Unacceptable ambiguity. *Linguistic Inquiry* 5: 17–68.

Hayes, Bruce P. 1997a. Four rules of inference for ranking argumentation. Unpubl. ms., Department of Linguistics, University of California, Los Angeles.

—. 1997b. Gradient well-formedness in Optimality Theory. Unpubl. handout, Department of Linguistics, University of California, Los Angeles.

—. 2000. Gradient well-formedness in Optimality Theory. In Dekkers et al. 2000, 88–120.

Hayes, Bruce P., and Margaret MacEachern. 1998. Folk verse form in English. *Language* 74(3): 473–507.

Hays, William L. 1964. *Statistics for Psychologists*. New York: Holt, Rinehart and Winston.

Hewson, Claire M., Dianna Laurent, and Carl M. Vogel. 1996. Proper methodologies for psychological and sociological studies conducted via the Internet. *Behavior Research Methods, Instruments, and Computers* 28: 186–191.

Höhle, Tilman N. 1982. Explikationen für “normale Betonung” und “normale Wortstellung”. In Werner Abraham, ed., *Satzglieder im Deutschen: Vorschläge zur syntaktischen, semantischen und pragmatischen Fundierung*, 75–153. Tübingen: Gunter Narr.

Jackendoff, Ray S. 1971. Gapping and related rules. *Linguistic Inquiry* 2: 21–35.

Jacobs, Joachim. 1988. Probleme der freien Wortstellung im Deutschen. In Inger Rosengren, ed., *Sprache und Pragmatik*, vol. 5 of *Working Papers*, 8–37. Department of German, Lund

University.

Johnson-Laird, P. N., and Favien Savary. 1999. Illusory inference: A novel class of erroneous deductions. *Cognition* 71(3): 191–229.

Kas, Mark. 1991. A voyage to the *wh*-Islands. In Mark Kas, Eric Reuland, and Co Vet, eds., *Language and Cognition: Yearbook of the Research Group for Linguistic Theory and Knowledge Representation*, vol. 1, 169–183. University of Groningen.

Katz, Jerrold J. 1964. Semi-sentences. In Fodor and Katz 1964, 400–416.

Keller, Frank. 1996a. *Extraction from Complex Noun Phrases: A Case Study in Graded Grammaticality*. Master's thesis, University of Stuttgart.

—. 1996b. How do humans deal with ungrammatical input? Experimental evidence and computational modelling. In Dafydd Gibbon, ed., *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld, October 1996*, 27–34. Berlin: Mouton de Gruyter.

—. 1997. Extraction, gradedness, and optimality. In Alexis Dimitriadis, Laura Siegel, Clarissa Surek-Clark, and Alexander Williams, eds., *Proceedings of the 21st Annual Penn Linguistics Colloquium*, no. 4.2 in Penn Working Papers in Linguistics, 169–186. Department of Linguistics, University of Pennsylvania.

—. 1998. Gradient grammaticality as an effect of selective constraint re-ranking. In Gruber, Higgins, Olson, and Wysocki 1998, 95–109.

—. 1999. Book review: The empirical base of linguistics: Grammaticality judgments and linguistic methodology, Carson T. Schütze. *Journal of Logic, Language and Information* 8(1): 114–121.

—. 2000. Evaluating competition-based models of word order. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 747–752. Mahwah, NJ: Lawrence Erlbaum Associates.

—. 2001. Experimental evidence for constraint competition in gapping constructions. In Gereon Müller and Wolfgang Sternefeld, eds., *Competition in Syntax*. Berlin: Mouton de Gruyter.

Keller, Frank, and Theodora Alexopoulou. 2001. Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, in press.

Keller, Frank, and Ash Asudeh. 2000. Probabilistic learning algorithms and Optimality Theory. Submitted.

Keller, Frank, Martin Corley, Steffan Corley, Lars Konieczny, and Amalia Todirascu. 1998.

WebExp: A Java toolbox for web-based psychological experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

Keller, Frank, and Antonella Sorace. 2000. Gradient auxiliary selection and impersonal passivization in German: An experimental investigation. Submitted.

Kim, Albert. 1994. Graded unification: A framework for interactive processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Las Cruces, NM. Student Session.

Kiss, Katalin É., ed. 1995. *Discourse Configurational Languages*. Oxford: Oxford University Press.

Kluender, Robert. 1992. Deriving island constraints from principles of predication. In Helen Goodluck and Michael Rochemont, eds., *Island Constraints: Theory, Acquisition and Processing*, 223–258. Dordrecht: Kluwer.

Kuno, Susumo. 1976. Gapping: A functional analysis. *Linguistic Inquiry* 7: 300–318.

—. 1987. *Functional Syntax: Anaphora, Discourse and Empathy*. Chicago: University of Chicago Press.

Kwasny, Stan C., and Norman K. Sondheimer. 1979. Ungrammaticality and extragrammaticality in natural language understanding systems. In *Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics*, 19–23. La Jolla, CA.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4): 715–762.

Ladd, Robert D. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Lakoff, George. 1973. Fuzzy grammar and the performance/competence terminology game. In Claudia Corum, T. Cedric Smith-Stark, and Ann Weiser, eds., *Papers from the 9th Meeting of the Chicago Linguistic Society*, 271–291. Chicago.

Lapata, Maria, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 30–36. Bergen.

Lasnik, Howard, and Mamoru Saito. 1984. On the nature of proper government. *Linguistic Inquiry* 15(2): 235–289.

Lee, Hanjung. 1998. Discourse competing with syntax: Prominence and “misplaced” QUE in child French. Unpubl. handout, Stanford University.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990a. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: An application. In COGSCI 1990, 884–891.

—. 1990b. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In COGSCI 1990, 388–395.

—. 1991. Unifying syntactic and semantic approaches to unaccusativity: A connectionist approach. In *Proceedings of the 17th Annual Meeting of the Berkeley Linguistics Society*, 388–395. Berkeley.

Legendre, Géraldine, Anne Vainikka, Marina Todorova, and Paul Hagstrom. 1998. Optional finiteness in early child grammars. Unpubl. handout, Department of Cognitive Science, Johns Hopkins University, Baltimore.

Legendre, Géraldine, Colin Wilson, Paul Smolensky, Kristin Homer, and William Raymond. 1995. Optimality and *Wh*-Extraction. In Jill Beckman, Laura Walsh Dickey, and Suzanne Urbanczyk, eds., *Papers in Optimality Theory*, no. 18 in Occasional Papers in Linguistics, 607–636. University of Massachusetts, Amherst.

Lenerz, Jürgen. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Stauffenburg.

Levin, Beth, and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge, MA: MIT Press.

Lodge, Milton. 1981. *Magnitude Scaling: Quantitative Measurement of Opinions*. Beverley Hills, CA: Sage Publications.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marks, Lawrence E. 1967. Judgments of grammaticalness of some English sentences and semi-sentences. *American Journal of Psychology* 20: 196–204.

—. 1968. Scaling of grammaticalness of self-embedded English sentences. *Journal of Verbal Learning and Verbal Behavior* 7: 965–967.

McDaniel, Dana, and Wayne Cowart. 1999. Experimental evidence of a minimalist account of English resumptive pronouns. *Cognition* 70: B15–B24.

McDonald, Scott. 1995. *Learning Compound Order: Towards a Functional Explanation*. Master's thesis, University of Edinburgh.

Mehler, Jacques. 1999. Editorial. *Cognition* 71(3): 187–189.

Meng, Michael, Markus Bader, and Josef Bayer. 1999. Sprachverstehem im Kontext: Zur Rolle struktureller und semantisch-pragmatischer Information bei der Auflösung von Subjekt-Objekt-Ambiguitäten im Deutschen. Unpubl. ms., Institute for Linguistics, University of Jena.

Meyer, Charles F. 1979. The greater acceptability of certain English elliptical coordinations.

Studia Linguistica 33(2): 130–137.

Mitchell, Tom. M. 1997. *Machine Learning*. New York: McGraw-Hill.

Mohan, B. A. 1977. Acceptability testing and fuzzy grammar. In Sidney Greenbaum, ed., *Acceptability in Language*, 133–148. The Hague: Mouton.

Mohanan, K. P. 1993. Fields of attraction in phonology. In John A. Goldsmith, ed., *The Last Phonological Rule: Reflections on Constraints and Derivations*. Chicago: University of Chicago Press.

Müller, Gereon. 1999. Optimality, markedness, and word order in German. *Linguistics* 37(5): 777–818.

Nagata, Hiroshi. 1987. Long-term effects of repetition on judgments of grammaticality. *Perceptual and Motor Skills* 65(1): 265–299.

—. 1988. The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *Journal of Psycholinguistic Research* 17(1): 1–17.

—. 1989a. Effect of repetition on grammaticality judgments under objective and subjective self-awareness conditions. *Journal of Psycholinguistic Research* 18(3): 255–269.

—. 1989b. Judgments of sentence grammaticality and field-dependence of subjects. *Perceptual and Motor Skills* 69(3): 739–747.

—. 1989c. Repetition effects in judgments of grammaticality of sentences: Examination with ungrammatical sentences. *Perceptual and Motor Skills* 68(1): 275–282.

Nagy, Naomi, and Bill Reynolds. 1997. Optimality Theory and variable word-final deletion in Faetar. *Language Variation and Change* 9: 37–55.

Pafel, Jürgen. 1998. *Skopus und logische Struktur: Studien zum Quantorenskopus im Deutschen*. Habilitationsschrift, University of Tübingen.

Papp, Szilvia. 2000. Stable and developmental optionality in native and non-native Hungarian grammars. *Second Language Research* 16(2): 173–200.

Pechmann, Thomas, Hans Uszkoreit, Johannes Engelkamp, and Dieter Zerbst. 1994. Word order in the German middle field: Linguistic theory and psycholinguistic evidence. CLAUS Report 43, Department of Computational Linguistics, Saarland University.

Pesetsky, David. 1987. *Wh*-in-situ: Movement and unselective binding. In Eric J. Reuland and Alice G. B. ter Meulen, eds., *The Representation of (In)definiteness*, 98–129. Cambridge, MA: MIT Press.

—. 1998. Some optimality principles of sentence pronunciation. In Barbosa et al. 1998, 337–383.

Philippaki-Warburton, Irene. 1985. Word order in Modern Greek. *Transactions of the Philological Society* 113–143.

Pierrehumbert, Janet B., and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, eds., *Intentions in communication*, 271–311. MIT Press, Cambridge.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Prévost, Philippe, and Lydia White. 2000. Missing surface inflection or impairment in second language acquisition? evidence from tense and agreement. *Second Language Research* 16(2): 103–133.

Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical Report 2, Center for Cognitive Science, Rutgers University.

—. 1997. Optimality: From neural networks to universal grammar. *Science* 275: 1604–1610.

Prince, Ellen F. 1986. On the syntactic marking of presupposed open propositions. In A. Farley, P. Farley, and K.-E. McCullough, eds., *Papers from the 22nd Meeting of the Chicago Linguistic Society*, vol. 2: The Parasession on Pragmatics and Grammatical Theory, 208–222. Chicago.

Quirk, Randolph. 1965. Descriptive statement and serial relationship. *Language* 41(2): 205–217.

Reinhart, Tanya. 1982. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27(1): 53–94.

Reinhart, Tanya, and Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry* 24(4): 657–720.

Reynolds, Bill. 1994. *Variation and Phonological Theory*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Rietveld, Toni, and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin: Mouton de Gruyter.

Riezler, Stefan. 1996. Quantitative constraint logic programming for weighted grammar applications. In *Proceedings of the 1st Conference on Logical Aspects of Computational Linguistics*. Berlin: Springer.

—. 1998. *Probabilistic Constraint Logic Programming*. Ph.D. thesis, University of Tübingen.

Robertson, Dan. 2000. Variability in the use of the English article system by Chinese learners of English. *Second Language Research* 16(2): 135–172.

Robinson, Jane J. 1982. DIAGRAM: A grammar for dialogues. *Communications of the ACM* 25(1): 27–47.

Rochemont, Michael S. 1986. *Focus in Generative Grammar*. Amsterdam: John Benjamins.

Ross, John R. 1970. Gapping and the order of constituents. In Manfred Bierwisch and Karl Erich Heidolph, eds., *Progress in Linguistics: A Collection of Papers*, 249–259. The Hague: Mouton.

—. 1972. The category squish: Endstation hauptwort. In Paul M. Peranteau, Judith N. Levi, and Gloria C. Phares, eds., *Papers from the 8th Meeting of the Chicago Linguistic Society*, 316–338. Chicago.

—. 1973a. A fake NP squish. In Charles James N. Bailey and Roger W. Shuy, eds., *New Ways of Analyzing Variation in English*, 96–140. Washington: Georgetown University Press.

—. 1973b. Nouniness. In Osamu Fujimura, ed., *Three Dimensions of Linguistic Theory*, 137–257. Tokyo: TEC.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1: Foundations, 318–362. Cambridge, MA: MIT Press.

Sadock, Jerrold M. 1998. Grammatical tension. In Gruber et al. 1998, 179–198.

Samek-Lodovici, Vieri. 1996. *Constraints on Subjects: An Optimality Theoretic Analysis*. Ph.D. thesis, Rutgers University, Piscataway, NJ.

Sarle, Warren S. 1994. Neural networks and statistical models. In *Proceedings of the 19th Annual SAS Users Group International Conference*, 1538–1550. SAS Institute, Cary, NC.

Scheepers, Christoph. 1997. *Menschliche Satzverarbeitung: Syntaktische und thematische Aspekte der Wortstellung im Deutschen*. Ph.D. thesis, University of Freiburg.

Schneider-Zioga, P. 1994. *The Syntax of Clitic Doubling in Modern Greek*. Ph.D. thesis, University of Southern California, Los Angeles.

Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Seibert, Anja J. 1993. Intransitive constructions in German and the ergative hypothesis. Working Papers in Linguistics 14, Department of Linguistics, University of Trondheim.

Selkirk, Elisabeth. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Smolensky, Paul, Géraldine Legendre, and Yoshiro Miyata. 1992. Principles for an integrated connectionist/symbolic theory of higher cognition. Report CU-CS-600-92, Computer Science Department, University of Colorado at Boulder.

—. 1993. Integrating connectionist and symbolic computation for the theory of language.

Current Science 64: 381–391.

Sorace, Antonella. 1992. *Lexical Conditions on Syntactic Knowledge: Auxiliary Selection in Native and Non-Native Grammars of Italian*. Ph.D. thesis, University of Edinburgh.

—. 1993a. Incomplete vs. divergent representations of unaccusativity in non-native grammars of Italian. *Second Language Research* 9: 22–47.

—. 1993b. Unaccusativity and auxiliary choice in non-native grammars of Italian and French: Asymmetries and predictable indeterminacy. *Journal of French Language Studies* 3: 71–93.

—. 1998. Near-nativeness, optionality and L1 attrition. In *Proceedings of the 12th International Symposium of Theoretical and Applied Linguistics*. Thessaloniki.

—. 1999. Differential effects of attrition in the L1 syntax of near-native L2 speakers. In S. Catherine Howell, Sarah A. Fish, and Thea Keith-Lucas, eds., *Proceedings of the 24th Annual Boston University Conference on Language Development*, 719–725. Somerville, MA: Cascadilla Press.

—. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76(4): 859–890.

Sorace, Antonella, and Michela Cennamo. 2000. Aspectual constraints on auxiliary choice in Paduan. Unpubl. ms., University of Edinburgh and University of Naples.

Sorace, Antonella, and Wietske Vonk. 1998. Gradient effects of unaccusativity in Dutch. Unpubl. ms., Department of Theoretical and Applied Linguistics, University of Edinburgh and Max Planck Institute for Psycholinguistics, Nijmegen.

Steedman, Mark. 1990. Gapping as constituent coordination. *Linguistics and Philosophy* 13: 207–264.

—. 1991. Structure and intonation. *Language* 67(2): 260–296.

Sternefeld, Wolfgang. 1998. Suboptimale syntaktische Strukturen im Deutschen. Project Proposal, SFB 441, University of Tübingen.

Stevens, S. S. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: John Wiley.

Stolz, Walter R. 1967. A study of the ability to decode grammatically novel sentences. *Journal of Verbal Learning and Verbal Behavior* 6: 867–873.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29(2): 229–268.

Tesar, Bruce B. 1998. Error-driven learning in Optimality Theory via the efficient computation of optimal forms. In Barbosa et al. 1998, 421–435.

Tsimpli, Maria Ianthi. 1995. Focusing in Modern Greek. In Kiss 1995, 176–206.

Tsiplakou, Stavroula. 1998. *Focus in Greek: Its Structure and Interpretation*. Ph.D. thesis, University of London, School of Oriental and African Studies.

Uszkoreit, Hans. 1987. *Word Order and Constituent Structure in German*. Stanford, CA: CSLI Publications.

Valiouli, Maria. 1994. Anaphora, agreement, and right dislocation in Greek. *Journal of Semantics* 11: 55–82.

Vallduví, Enric. 1992. *The Informational Component*. New York: Garland.

—. 1995. Structural properties of information packaging in Catalan. In Kiss 1995, 122–152.

Vallduví, Enric, and Elisabet Engdahl. 1996. The linguistic realization of information packaging. *Linguistics* 34(3): 459–519.

van Hout, Angeliek, Janet Randall, and Jürgen Weissenborn. 1993. Acquiring the unergative-unaccusative distinction. In Maaike Verrips and Frank Wijnen, eds., *The Acquisition of Dutch*, no. 60 in Amsterdam Series in Child Language Development. Institute for Linguistics, University of Amsterdam.

Vendler, Zeno. 1967. *Linguistics in Philosophy*. Ithaca, NY: Cornell University Press.

Vion, Monique, and Annie Colas. 1995. Contrastive marking in French dialogue: Why and how. *Journal of Psycholinguistic Research* 24(5): 313–331.

Warner, John, and Arnold L. Glass. 1987. Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *Journal of Memory and Language* 26: 714–738.

Weischedel, Ralph M., and John E. Black. 1980. Responding intelligently to unparsable inputs. *American Journal of Computational Linguistics* 6(2): 97–109.

Zaenen, Annie. 1993. Unaccusativity in Dutch: Integrating syntax and lexical semantics. In James Pustejovsky, ed., *Semantics and the Lexicon*, 129–161. Dordrecht: Kluwer.

Index of Citations