How Do Humans Deal with Ungrammatical Input? Experimental Evidence and Computational Modelling

Frank Keller
Institute for Computational Linguistics
University of Stuttgart
keller@ims.uni-stuttgart.de

Abstract

We show how psycholinguistic methods can be used to investigate degrees of grammaticality and report experimental results for gradedness in extraction from complex NPs. We argue for the theoretical relevance of psycholinguistic evidence and sketch a model for gradedness based on ranked constraints. Our approach builds on results from optimality theory and can be computationally implemented using weighted constraint-based grammars.

Wir demonstrieren die Verwendung von psycholinguistischen Methoden zur Untersuchung von Grammatikalitätsgraden und stellen experimentelle Ergebnisse zur Gradiertheit bei der Extraktion aus komplexen NPs vor. Wir argumentieren für die theoretische Relevanz psycholinguistischer Evidenz und skizzieren ein Modell für Grammatikalitätsgrade basierend auf priorisierten Constraints. Unser Ansatz baut auf Ergebnisse der Optimalitätstheorie auf und kann durch gewichtete Constraintgrammatiken implementiert werden.

1 Introduction

Not all ungrammatical sentences are equally bad. This observation dates back at least to Chomsky (1964), and on an informal level, degrees of (un-)grammaticality are regularly used to support claims in linguistic theory. Examples can be found in the GB textbook by Haegeman (1994), who cites the standard assumption that subjacency violations result in only mild ungrammaticality, while ECP violations cause strong ungrammaticality. Without attempting an explicit account of gradedness, she conjectures that "[g]iven that there are various principles of grammar which may be violated the sentence will worsen as more than one principle is violated" (Haegeman 1994:568).

Dafydd Gibbon (ed.). *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld, October 1996*, 27–34. Berlin: Mouton de Gruyter, 1996.

This seems to be a typical case: even though the existence of graded data and its potential significance to linguistic research is generally acknowledged, hardly any effort has gone into the theoretical investigation of graded grammaticality. On the empirical side, we lack proper criteria for the gathering of graded data, and the use and interpretation of intermediate grammaticality ratings varies greatly between researchers.

In this paper, we show how psycholinguistic methods can be used to obtain reliable graded data. To model graded grammaticality, we propose an extended version of Optimality Theory and sketch a computational implementation using weighted constraint-based grammars.

2 Phenomena

To provide a test case for the psycholinguistic study of gradedness, we conducted experiments on extraction from complex NPs, which are standardly assumed to be islands for extraction. Picture NPs constitute well-known counter-examples, as they allow for extraction in certain cases. Kluender (1992) claims that extractability from picture NPs depends on the specifier of the picture NP and observes that acceptability gradually decreases from (1a) to (1e):

(1)
a. Who did you see pictures of?
b. Who did you see *a* picture of?
c. Who did you see *the* picture of?
d. Who did you see *his* picture of?
e. Who did you see *John's* picture of?

Kluender (1992) attributes this hierarchy to the increase in specificity of the picture NP, and Fiengo (1987) notes that the definiteness and singularity contribute to specificity. Extractability also depends on the matrix verb: Kluender (1992) gives the following pairs, where the first one of the verbs is more acceptable than the second one:

(2)
a. What did John *have/analyze* a picture of?
b. What did John *see/criticize* a picture of?
c. What did John *find/discuss* a picture of?
d. What did John *draw/lose* a picture of?
e. What did John *develop/destroy* a picture of?

Another factor seems to be the specificity of the the extracted NP. Evidence comes from extraction from relative clause islands, which Kluender (1992) claims becomes worse with decreasing specificity, i.e., from (3a) to (3c):

(3)
a. *Which paper* do you really need to find someone you can intimidate with?

b. *How many papers* do you really need to find someone you can intimidate with?

c. *What* do you really need to find someone you can intimidate with?

3 Experimental Evidence

3.1 Method

Experimental data for this study were elicited using magnitude estimation (ME), an experimental technique standardly applied in psychophysics to measure judgements of sensory stimuli (cf. Stevens 1975). The ME procedure requires subjects to estimate the magnitude of physical stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. Highly reliable judgements can be achieved in this way for a whole range of sensory modalities, such as brightness, loudness, or tactile stimulation.

The ME paradigm has been extended successfully to the psychosocial domain (cf. Lodge 1981) and recently Bard et al. (1996) showed that linguistic judgements can be elicited in the same way as judgements for sensory or social stimuli. In contrast to the 5- or 7-point scale conventionally used to measure psychological intuitions, ME employs a continuous numerical scale, thus providing fine-grained measurements of linguistic acceptability, which are robust enough to yield statistically significant results, while being highly replicable both within and across speakers. In psycholinguistics, ME has been applied to phenomena such as unaccusativity (cf. Sorace 1993a), auxiliary selection (cf. Sorace 1993b), and compounding (cf. McDonald 1995), both in native speakers and in second language learners.

ME requires subjects to assign numbers to a series of linguistic stimuli proportional to the acceptability they perceive. First, subjects are exposed to a modulus item, which they assign an arbitrary number. Then, all other stimuli are rated proportional to the modulus, i.e., if a sentence is three times as acceptable as the modulus, it gets three times the modulus number, etc. In the study described below, stimuli were presented on a computer screen, and subjects had to respond by keying in numerical acceptability judgements. Display time was limited to 6000 ms in order to elicit immediate judgements, leaving no time for metalinguistic reflections.

3.2 Test Corpus

Subjects had to judge sentences from a test corpus containing examples for extraction from picture NPs. The corpus was designed to investigate the following linguistic factors:

- *Def:* definiteness of the picture NP; indefinite vs. definite determiner

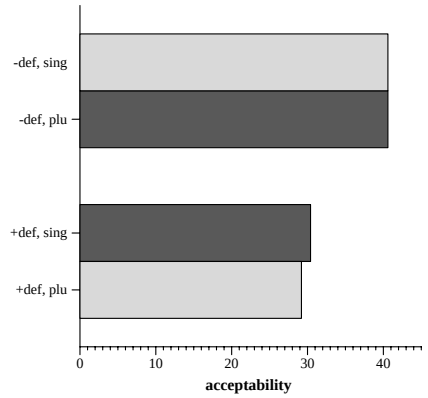

Figure 1: Mean acceptability vs. definiteness/number of picture NP

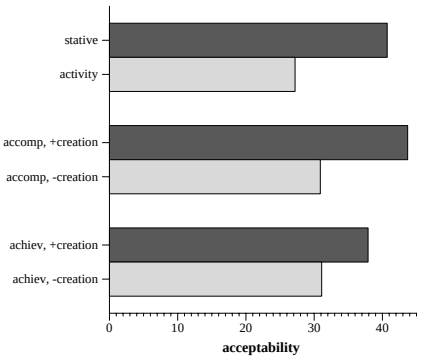

Figure 2: Mean acceptability vs. semantic class of matrix verb

- *Num:* number of the picture NP; plural vs. singular
- *Asp:* aspectual class of the matrix verb; active, stative, achievement, accomplishment verb; +/− creation verb
- *Spec:* specificity of the extracted NP; *which N*, *how many N*, *who*, *what*

We used an experimental design of $2 \times 2 \times 6 \times 4$ factorial (Def $\times$ Num $\times$ Asp $\times$ Spec), yielding a total of 96 conditions. The test corpus contained examples analogous to the ones in (1)–(3), with varying lexicalizations, providing a total of 576 tokens. Nineteen native speakers of English participated in the study, all of them naive (i.e., non-linguists). The stimuli were placed in a latin square design and each subject had to rate 96 test and 108 filler sentences.

3.3 Results

Mean acceptability is graphed in fig. 1 for the factors Def and Num, in fig. 2 for Asp, and in fig. 3 for Spec. The results of an Anova carried out on the log-transformed data are given in fig. 4. Post-hoc tests show significant differences for all relevant conditions (cf. Keller 1996 for details).

The significant differences in acceptability for the factors Def and Asp confirm the predictions by Kluender (1992). The factor Num failed to be significant, in contrast the assumption of Fiengo (1987). For Spec, an acceptability hierarchy was found, however different from the one claimed by Kluender (1992). This suggests that ME data can yield results that are either compatible or incompatible with theoretical claims, and hence can be used as a tool for evaluating predictions from linguistic theories, which are normally based on purely intuitive evidence.

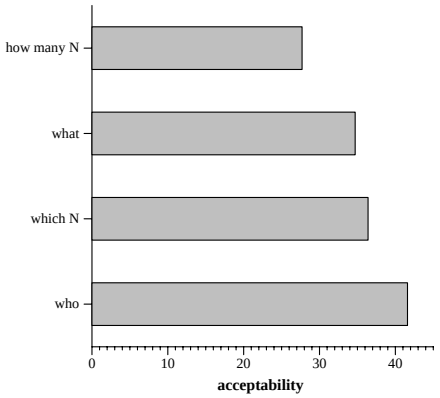

Figure 3: Mean acceptability vs. specificity of extracted NP

Main effects	F	p
Def	9.119	0.007
Spec	8.701	0.000
Num	0.319	0.579
Asp	8.568	0.000

Interactions	F	p
Def/Spec	2.919	0.042
Spec/Num	5.200	0.003
Spec/Asp	2.359	0.003

Figure 4: Anova results (only significant interactions listed)

4 Computational Modelling

The only computational model of graded grammaticality that we are aware of is the one by Erbach (1995). In this framework, numerical preference values are attached to the constraints a grammar (expressed by the # operator), and the overall preference of a clause is computed as the weighted sum of the preferences of the constraints it satisfies:

(4) $$C\#(p \cdot (w_1 \cdot p_1 + \ldots + w_n \cdot p_n)) \leftarrow A_1\#p_1 \wedge \ldots \wedge A_n\#p_n.$$

Here, C is a clause with preference p that has to satisfy the constraints $A_1, \ldots, A_n$ with preferences $p_1, \ldots, p_n$ and weights $w_1, \ldots, w_n$. Erbach (1995) assumes a probabilistic distribution for the preferences, and claims that the probability his model assigns to a linguistic structure corresponds to its degree of grammaticality. Although we do not share this assumption,[1] we conjecture that his approach can be used to implement a numerical model of gradedness based on ME data as presented in sec. 3. However, two principal problems arise:

(a) Per se, such a numerical model is underdetermined. If arbitrary numerical information is allowed on arbitrary clauses, the grammar fails to make interesting theoretical predictions, as it can be set up to accommodate any data. It is unclear what would constitute counter-evidence to a linguistic theory couched as a numerical grammar.

(b) No suitable method is available for determining the weights and probabilities of such a numerical model. Until now, no algorithms to train these values from corpus data exist for models as powerful as the one of Erbach (1995),[2] and just

[1] For a detailed discussion, cf. Keller 1996.

[2] Erbach (1995) hypothesizes that the EM algorithm could be used to train the probabilities, but Riezler (1996) shows that this leads to serious problems with reentrancies.

stipulating them seems theoretically unsound.[3]

To solve these problems, we start from the assumption that gradedness in linguistic data is due to the fact that certain grammatical constraints are "stronger" than others and hence lead to a greater degree of ill-formedness when violated. We propose a linguistic model where constraints are ranked for strength and the well-formedness of a structure is computed from the ranks of the constraints it violates.

Ranked constraints are standardly employed by Optimality Theory (OT, cf. Grimshaw 1995), a declarative variant of GB building on the following basic assumptions: (a) Constraints can be violated. (b) Constraints are hierarchically ordered. (c) In all languages, the same constraints apply. Crosslinguistic variation is due to variation in the constraint hierarchy (re-ranking). (d) The grammatical structure of an utterance is determined as the optimal analysis from a set of possible candidates, where "optimal" is defined as satisfying most of the most highly ranked constraints.

Standard OT treats all suboptimal candidates as equally ungrammatical, which leads to a binary notion of grammaticality. We propose to extend OT so as to assign degrees of grammaticality to competing candidates according to the ranks of the constraints they violate. Such a framework then can be used to model graded data as presented in sec. 3.

To overcome problem (a), we can make use of OT's well-established and empirically validated assumptions on how constraint rankings are to be devised, interpreted, and tested. If we set up the ranks in our grammar in accordance with OT assumptions, we avoid underdeterminedness and obtain a model that makes testable predictions. Note that ranks in OT are expressed as a partial order on constrains, hence no commitment to actual constraint weights has to be made when the ranks are determined.

An extended OT grammar then predicts that a certain structure is more or less ungrammatical than another one (as it violates more or less highly ranked constraints), a prediction that can be tested against ME data. The following interesting consequences ensue:

- Under the assumption that suboptimal candidates are not equally ungrammatical, graded judgements can provide evidence for constraint rankings in OT grammars.
- If crosslinguistic variation is really due to constraint re-ranking (assumption (c)), then we expect crosslinguistic differences in the degree to which a certain construction is grammatical in a given language.

As for problem (b), our proposal entails that the values for the numerical component of a grammar do not need to be stipulated, but can be calculated directly from the optimality theoretic ranks of its constraints. The calculation scheme derives from the OT assumption that ranks are absolute: for a structure S let

[3]Note that Erbach (1995) does indeed stipulate the probabilities and weights for his model of German word order.

r be the rank of the most highly ranked constraint S violates. Then S is less optimal than any structure S' with r' lower than r, where r' is the rank of the most highly ranked constraint S' violates. This entails that the violation of a constraint of rank r cannot be compensated by the satisfaction of constraints ranked lower than r. Hence we need an exponential scheme to compute weights from ranks: given a rank $r_i \in \{0, \ldots, n-1\}$ for a constraint A_i in clause (4), the corresponding weight is $w_i = 2^{r_i}/(2^n - 1)$. For the preferences $p, p_1, \ldots, p_n$, the appropriate assumption is that a constraint yields the preference 1 if it is satisfied, and 0 otherwise. This implements the absoluteness of ranks in OT.

References

Bard, Ellen G., Dan Robertson, and Antonella Sorace. 1996. Magnitude Estimation of Linguistic Acceptability. *Language* 72: 32–68.

Chomsky, Noam. 1964. Degrees of Grammaticalness. In Jerry A. Fodor and Jerrold J. Katz, eds., *The Structure of Language: Readings in the Philosophy of Language*, 384–389. Englewood Cliffs, NJ: Prentice-Hall.

Erbach, Gregor. 1995. Bottom-Up Earley Deduction for Preference-Driven Natural Language Processing. Ph.D. thesis, University of the Saarland. Draft of August 31, 1995.

Fiengo, Robert. 1987. Definiteness, Specificity, and Familiarity. *Linguistic Inquiry* 18: 163–166.

Grimshaw, Jane. 1995. Projection, Heads, and Optimality. Unpubl. ms., Department of Linguistics and Center for Cognitive Science, Rutgers University.

Haegeman, Liliane. 1994. *Introduction to Government and Binding Theory*. Oxford: Basil Blackwell, 2nd edn.

Keller, Frank. 1996. Extraction from Complex Noun Phrases. A Case Study in Graded Grammaticality. Master's thesis, Institute for Computational Linguistics, University of Stuttgart.

Kluender, Robert. 1992. Deriving Island Constraints from Principles of Predication. In Helen Goodluck and Michael Rochemont, eds., *Island Contraints: Theory, Acquisition and Processing*, 223–258. Dordrecht: Kluwer.

Lodge, Milton. 1981. *Magnitude Scaling: Quantitative Measurement of Opinions*. Beverley Hills, CA: Sage Publications.

McDonald, Scott. 1995. Learning Compound Order: Towards a Functional Explanation. Master's thesis, Centre for Cognitive Science, University of Edinburgh.

Riezler, Stefan. 1996. Quantitative Extensions of Constraint Logic Grammars. Unpubl. ms., University of Tübingen.

Sorace, Antonella. 1993a. Incomplete vs. Divergent Representations of Unaccusativity in Non-Native Grammars of Italian. *Second Language Research* 9: 22–47.

Sorace, Antonella. 1993b. Unaccusativity and Auxiliary Choice in Non-Native Gram-

mars of Italian and French: Asymmetries and Predicable Indeterminacy. *Journal of French Studies* 3: 71–93.

Stevens, S. S. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: John Wiley.