Talkers account for listener and channel characteristics to communicate efficiently

John K Pate[a,*], Sharon Goldwater[b]

[a]Present address: Department of Computing, Macquarie University, Sydney, NSW 2109, Australia (+61) 414 796 255
[b]School of Informatics, The University of Edinburgh, 10 Crichton St, Edinburgh, UK, EH8 9AB

Abstract

A well-known effect in speech production is that more predictable linguistic constructions tend to be reduced. Recent work has interpreted this effect in an information-theoretic framework, proposing that such *predictability effects* reflect a tendency towards communicative efficiency. However, others have argued that these effects are, in the terminology of Gould and Lewontin (1979), *spandrels*: incidental by-products of other processes (such as a talker-oriented tendency for low production effort). This article develops the information-theoretic framing more fully, showing that information-theoretic efficiency involves different kinds of coding operations (predictability effects), not all of which are consistent with the spandrel account. Using mixed effects regressions, we analyze word durations in several spontaneous speech corpora, comparing predictability effects between infant-directed and adult-directed speech and between speech to visible and invisible listeners. We find that talkers adjust the extent to which production varies with predictability measures according to listener characteristics, and exploit an additional visual channel to eliminate phonetic redundancy. This pattern would demand multiple independent spandrel accounts, but is unified by an adaptive account. Our results broaden the scope of existing work on predictability effects and provide further evidence that these effects are tied to communicative efficiency.

1. Introduction

Talkers can usually communicate a given message using many different linguistic forms: for a fixed meaning or intention, talkers can produce different syntactic constructions, select different words with similar meanings, and pronounce parts of the utterance more or less clearly. Some of these productions will be longer and more distinct (overarticulated), while others will be shorter and more ambiguous (reduced). For example,

*Corresponding Author
Email addresses: `john.pate@mq.edu.au` (John K Pate), `sgwater@inf.ed.ac.uk` (Sharon Goldwater)

the pronunciation of *feline* typically takes longer than a pronunciation of *cat*, but the pronunciation of *cat* is more ambiguous (e.g. its phonological form is a substring of many other words).

Many researchers have argued that talkers make these choices, whether consciously or unconsciously, in ways that lead to more efficient communication. Lindblom (1990), focusing on word production, observed that speech appeared to vary on virtually every acoustic dimension according to a range of variables, including local predictability, talker identity, listener needs, and so forth. Lindblom proposed that speech lacked reliable invariants because the goal of speech is not the approximation of some ideal, but discrimination among different items in a lexicon. His **H**yper- & **H**ypo-articulation (H&H) theory proposed that talkers articulate "just enough," in terms of effort or acoustic distinctiveness, to enable reliable discrimination. Words that can be discriminated from other possible words on the basis of non-acoustic grounds require less acoustic distinctiveness, and so less articulatory effort, while talkers must provide more acoustic information for words that are harder to identify.

Subsequent proposals (e.g. Aylett and Turk, 2004; Jaeger, 2010) have sought to elaborate what "just enough" means in different ways, which we will discuss shortly, but they all similarly hinge on the observation that talkers tend to produce more reduced (more ambiguous) forms for portions of their message that are more probable. Such *predictability effects* are pervasive throughout language, resulting in shorter phonological forms for more common words (Zipf, 1949), phonetic reduction of words that are discourse given or have higher *n*-gram probabilities (Aylett and Turk, 2004; Bell et al., 2009; Seyfarth, 2014; Priva, 2008), and reduction or omission of words that have higher syntactic probabilities (Levy and Jaeger, 2007; Frank and Jaeger, 2008; Tily et al., 2009; Gahl and Garnsey, 2004; Gahl et al., 2006). Broadly, predictability effects are information-theoretically efficient because they make shorter utterances more common than longer ones while guarding against communication errors.

Although previous work has shown that predictability effects exist, and that they make speech more efficient, it has not been established that predictability effects are an adaptation for efficient communication for two reasons. First, it is difficult to measure *how much* more efficient communication is as a result of predictability effects, and so we do not know if they have any practical impact. Second, it is possible to derive at least some predictability effects as a consequence of how lexical and grammatical knowledge might be stored and accessed. For example, Dell and Brown (1991) and Ferreira (2003) have highlighted the possibility that some aspects of linguistic knowledge and processing are shared between production and comprehension. With this kind of sharing, the same items that are hard to access in comprehension will be hard to access for production. If there is also some mechanism that lengthens the realization of hard-to-access items, talkers would end up producing longer realizations for items that would be hard for them to comprehend, if they were the listener. However, this situation would not necessarily lead to a practical improvement in communication rate or reliability: the magnitude of this effect is not tuned for efficiency but is simply a consequence of whatever system lengthens the realization of hard-to-access items. Thus, if predictability effects make communication only negligibly more efficient, and result from architectural accidents, they cannot be said to reflect adaptation to communicative efficiency in any real sense.

This example unearths the worry that predictability effects may be what Gould and Lewontin (1979) term "evolutionary spandrels," by analogy with a particular architectural feature of buildings with arches. Gould and Lewontin observed that the most striking features of arches are the designs and mosaics that appear in the roughly triangular region, called the spandrel, between arches at right angles. Despite the striking appearance, the spandrels are not a driving feature of the architecture: they exist because arches must be curved to support weight. Gould and Lewontin (1979) used this example to illustrate the point that even very striking biological features may be evolutionary coincidences or by-products, without being adaptive themselves. A general correlation between distinctiveness, or ease of perception, and coarse measures of predictability is not sufficient to establish that language is adapted for communicative efficiency, no matter how tantalizing the prospect.

One especially plausible spandrel account appeals to production or planning difficulty: people avoid productions that require more effort, and these same productions tend to be rare. Under this account, production difficulty alone is the driving force, and there should not be an independent effect of predictability. However, recent experimental work has found an independent contribution of predictability, after controlling for production difficulty (Jaeger, 2013). For example, Kurumada and Jaeger (2013) examined the production of optional case markers in Japanese that are sometimes redundant with other grammatical function cues, and found independent contributions of both production difficulty and predictability. Baese-Berk and Goldrick (2009), with replications using different methodology by Peramunage et al. (2011), Kirov and Wilson (2012), and Buz et al. (2014), looked at voice-onset time (VOT), an important cue to voicing contrasts in English stops. They found that talkers produce a stronger VOT contrast when both words of a voicing minimal pair have plausible referents in the current discourse context. Buz et al. (2014) additionally looked for an effect of planning difficulty (as measured by production latency), but did not find one. Although this last result, as all null results, should be interpreted with caution, these studies together begin to provide evidence that predictability effects on at least morphology and phonetic detail cannot be solely attributed to production difficulty.

This paper provides a complementary line of evidence favoring the view that predictability effects reflect an adaptation towards communication. We investigate how distinctiveness correlates with different sources of predictability. These different sources are unified under an account that appeals to information-theoretic efficiency, but, by virtue of their differences, would presumably require independent spandrel accounts. To do this, we first develop the information-theoretic framework of recent work on predictability effects (Aylett and Turk, 2004; Levy and Jaeger, 2007; Jaeger, 2010) to classify predictability effects into three kinds: source coding from the talker's perspective, source and channel coding specialized to listener characteristics, and channel coding. These three kinds of predictability effects consider very different features, and so would presumably involve independent spandrel accounts. They are unified, however, by their relevance to information-theoretic optimization. With this theoretical framework in mind, we provide two corpus studies that find evidence for the latter two kinds of predictability effects. Concretely, we use mixed-effects regression models to analyze several corpora of spontaneous speech, examining predictability measures that have been previously shown to correlate with changes in word duration (i.e., time between word onset and offset) (Aylett and Turk, 2004; Bell et al., 2009; Gahl et al., 2012).

Our first analysis examines the relationship between these predictability measures and word duration in speech directed to prelinguistic infants, as compared to speech directed to adults. We find, first, that many of the patterns previously found in adult-directed speech also occur in infant-directed speech; this is, as far as we know, the first analysis to find evidence of predictability effects in infant-directed speech. More importantly, however, we also find a difference between adult-directed and infant-directed speech in how predictability (as measured by contextual probability) affects word duration. This difference suggests that talkers adjust predictability effects in response to listener characteristics.

Our second analysis compares predictability effects in speech directed to a visible listener and speech directed to an invisible listener. Again, we find different effects of predictability in the two conditions: specifically, a significantly weaker effect of unigram word probability on duration in the visible condition. This difference is in the direction predicted by an information-theoretic efficiency strategy, where talkers exploit the extra communicative capacity afforded by the visual channel. Together, these results suggest that the effects of predictability on speech production differ according to communicative demands, both in response to listener characteristics and to exploit increased channel capacity. These findings provide further evidence that predictability effects are not merely a ‘spandrel’ of language processing or general cognition, but are an important part of a system adapted towards efficient communication in the information-theoretic sense (Aylett and Turk, 2004; Levy and Jaeger, 2007; Jaeger, 2006).

It is important to note that we are making no claims about whether the differences in predictability effects we see in different communicative scenarios are due to conscious behavior on the part of the talker, or unconscious behavior. An effect can be deliberate without having a practical impact on communication. Conversely, an effect can be automatic, but have arisen because of pressure towards communicative efficiency, whether over the course of evolutionary time, language acquisition, or adaptation to a particular listener. How effects are implemented and their availability to conscious awareness and deliberate manipulation is a different question from their functional role and whether they exist due to pressure for efficient communication.

Before presenting our regression studies, we first review the information-theoretic foundations of efficiency accounts and how these relate to the existing literature on predictability effects and adaptation towards the listener and the communicative channel.

2. Background

Predictability effects have attracted attention because they reflect correlations that characterize information-theoretically efficient systems. In this section, we further develop the theoretical framing to provide one way of distinguishing between an adaptation account and a spandrel account. Efficient codes involve different coding operations (*source* and *channel*-coding) that have different functional roles (*brevity* and *error-avoidance*) and are sensitive to very different aspects of the current communicative situation. Because they are sensitive to such different kinds of information, if these different types of predictability effects exist but are spandrels, they would have to be independent spandrels. Thus, if such predictability effects exist, then the non-adaptive hypothesis would require multiple coincidences, making it less plausible than the hypothesis that the predictability effects are unified by an adaptation towards efficient communication.

2.1. Information Theory

Previous efficiency accounts have appealed to the notion of efficiency in the sense of the noisy channel theorem (Shannon, 1948) as an explanation for predictability effects (e.g. Aylett and Turk, 2004; Jaeger, 2010). This section outlines the relevant results from information theory, which formalizes communication as the endeavor by a sender, our talker, to communicate a *message* to a receiver, our listener. By hypothesis, the message cannot be sent directly to the receiver, so the sender *encodes* it into a different form, called the *signal*, that can be sent to the receiver. When analyzing linguistic communication, we may focus on many different message/signal pairs. For example, phonetics typically takes the signal to be something like a spectrogram, and the message to be variously segments, features, or gestures. Formal semantics, by contrast, typically takes the signal to be some syntactic analysis and the message to be meaning representations in first-order predicate logic.

From an information-theoretic point of view, a communication system is efficient to the extent that it satisfies two requirements. First, it should have, on average, short signals for messages. Second, it should also obtain a low probability of a communication error. To make this sense of efficiency concrete, and to see how it actually incorporates two kinds of encoding schemes, consider briefly a communication system whose messages are sequences of the 32 characters consisting of A-Z, space, and five punctuation symbols, and whose signals are sequences of binary digits (bits). One way to encode messages in this system is to impose an ordering on all of the characters (such as alphabetical ordering), and say that the signal for a character is a sequence of 1's equal to its position in the ordering followed by a 0 (so 'AB' is encoded '10110'). If the message characters occur with equal probability, this code uses an average of $\sum_{n=2}^{33} n\frac{1}{32} = \frac{560}{32} = 17.5$ signal characters per message character. Alternatively, we can notice that there are $2^5 = 32$ binary sequences of length 5, and choose to encode a message character with the five-digit binary number that indicates its position in the ordering. This second code uses only 5 signal characters per message character, and so is more efficient.

Now consider an 'ABC' communication system that still has binary signal characters but whose messages are just sequences of the characters 'A', 'B', and 'C'. We can unambiguously encode these three message characters with the signals '0', '10', and '11'; which message should be encoded with the short code? If the message characters occur with equal probability, it does not matter, but if one message character is more frequent than another, we can get shorter signal sequences, on average, by giving the short signal to the frequent character. Giving shorter signals to more common messages is called *source coding*, and works towards our first efficiency requirement of short signals. The minimum average number of signal characters per message character is called *entropy*, and a code that uses more signal characters per message character than the entropy is *redundant*. Effective source coding thus eliminates redundancy.

In particular, we can compute the entropy of a probability distribution over the set of possible messages

M as $H(M)$:

$$
\begin{aligned}
H(M) &= -\sum_{m \in M} P(m) \log(P(m)) \\
&= \sum_{m \in M} P(m) \log_b \left(\frac{1}{P(m)} \right)
\end{aligned} \tag{1}
$$

where b is the size of the signal alphabet. Equation 1 is the usual, concise presentation, but the version in the second line makes an important fact explicit: entropy is just a (weighted) average. The log term, called the Shannon information, is the length of the signal for each particular message m, using a signal alphabet with b characters, under the shortest possible code. The minimum average number of signal characters per message character is just the average of these lengths, weighted by their relative frequency $P(m)$. One strategy for good source coding (making a code shorter), then, would be to find messages whose Shannon information is substantially different from their signal length, and assign those messages a different code that is closer in length to their Shannon information.

However, up to this point we have not said anything about avoiding errors. Because source coding focuses on finding the shortest code that unambiguously identifies each message, there will be no errors if the receiver (listener) always correctly receives the signal that the sender (talker) transmits. However, such perfect transmission does not occur in real communicative scenarios, linguistic or otherwise, and the *noisy channel* model is an effort to reflect this imperfect transmission. It assumes that some of the signal characters are randomly changed or deleted during transmission, with these modifications called noise.[1] For example, for our 'ABC' code, a signal that is sent as '10' might be received as '11' or '00'. Thus, under the noisy channel, even if the talker produces a signal that unambiguously identifies a message, the signal that is received may identify a different message (or even many or no messages).

We can avoid errors due to noise by introducing extra signal characters that encode which signal character was intended. We might decide to reinforce the signal character '0' by following it with '11', and reinforce the signal character '1' by following it with '00', and taking a majority vote for the intended signal character. In this new code, the listener will recognize '011', '111', '010', and '001' all as the signal character '0', because they all have at least 2 bits of 3 that agree with the new code for '0'. Thus, this new code can tolerate one bit flip per message character without resulting in an error, and we could make it tolerate more bit flips by picking longer reinforcement codes. However, the code now produces longer signals without any change to

[1] If the changes are not random, for example '1111' always changes to '0000', then they are distortion rather than noise, and may be recovered by the listener by simply using a different mapping from messages to signals, or codebook.

the messages, meaning that it is redundant. Adding redundancy to the signal to avoid errors is called *channel coding*.

We might be worried that reliable channel coding is going to introduce so much redundancy to avoid errors that the nice short signals we got from the source coding become unmanageably long. The noisy channel theorem (Shannon, 1948) proves that, in fact, it is possible to get an error rate arbitrarily close to zero with a reasonably small amount of redundancy as long as we add in redundancy that anticipates the kind of errors the communication channel introduces. If we subtract this minimum amount of redundancy from the entropy of the source, we obtain the (fixed) channel capacity C which expresses the average length of the shortest code, per message character, that can be achieved with an arbitrarily low error rate (intuitively, we get the channel capacity by subtracting the amount of information that we must devote to reinforcing the signal from the amount of information we can effectively transmit).

To perform optimal source coding, talkers need to know the probability distribution over messages, and to perform optimal channel coding, they need to know the probability distribution over noise. In a practical situation, talkers could only ever have an estimate of these probability distributions. To the extent that their estimates are wrong, they will provide overly redundant messages.[2] Speech communication, then, can be made closer to optimal by considering more information in determining the probability of messages and noise.

With this information-theoretic framework, we can classify potential predictability effects into three kinds. First, some predictability effects could reflect source coding according to the general, overall probability of linguistic elements. Second, some predictability effects could reflect an effort to use knowledge about the listener to anticipate likely errors or messages. Third, some predictability effects could reflect an effort to specialize their channel probability distributions according to the rate or kinds of errors that can be expected under a given communicative scenario.

The considerations discussed so far address only communicative efficiency, and should be combined with a model of the cost of utterance planning for a more complete account of efficient language production strategies (Lindblom, 1990; Jaeger, 2013; Kurumada and Jaeger, To Appear). For example, Ferrer-i-Cancho (2005) presented an ideal talker model that linearly trades off maximizing the achieved communication rate with minimizing communication effort, operationalized as the entropy of the probability distribution

[2] Specifically, they will provide m more bits than necessary, where m is the Kullback-Leibler divergence of their estimated distributions from the true distributions.

over signals. By ignoring the cost of planning utterances, we here assume that the trade-off totally favors communication rate.

In the following subsections, we first review recent work on predictability effects within the information-theoretic framework, with an eye towards developing intuitions about these three types of effects. We then focus in on the latter two types of effects, which is where our own investigations will be directed. We discuss work from other traditions that can be re-interpreted through the lens of information theory, and the evidence this work provides regarding listener-based and channel-based predictability effects.

2.2. Source versus channel coding in predictability effects

Recent work on predictability effects has referred to information-theoretic principles in the abstract, but has not explicitly analyzed predictability effects in terms of different kinds of coding. Following the overview of coding theory above, we can understand a predictability effect as reflecting source coding to the extent that it *eliminates redundancy* for the purpose of *giving shorter signals to common messages*. Alternatively, we can understand a predictability effect as reflecting channel coding to the extent that it *preserves or introduces redundancy* for the purpose of *avoiding communication errors*.

One of the most well-known predictability effects in language is the inverse relationship between word frequency and word length (Zipf, 1949). Recently, Piantadosi et al. (2011) showed that in English, German, and Dutch, the average information content of a word (i.e., the average negative log probability given its *n*-gram context) predicts word length (in orthography, phonemes, and syllables) even better than raw frequency does; Manin (2006) found the same result for Russian. Piantadosi et al. argued that this result strongly supports the idea of a lexicon optimized for communicative efficiency. Seyfarth (2014) further found that talkers pronounce words with shorter duration (in seconds) if those words typically appear in highly predictable contexts. These authors did not explicitly discuss the type of coding their results suggest, but empirical results such as these, which relate the static forms of words to their probability of use, are clear examples of what one would expect from source coding: a lexicon that uses shorter signals for high-probability messages.

Other effects clearly reflect some degree of channel coding, because they introduce redundancy that is especially likely to prevent communication errors. In the Lombard effect, talkers involuntarily produce more intelligible speech when they hear noise (Lombard, 1911, 1910a,b). Such speech provides more acoustic redundancy, such as greater amplitude, slower speech, and more energy at higher frequencies, precisely when the received acoustic signal will be more ambiguous. Moreover, this speech is more intelligible than regular speech (Summers et al., 1988; Claude Junqua, 1996), while simply increasing vocal force, from whispering through to shouting, results in less intelligible speech (Pickett, 1956).

Remember that the information-theoretic notion of noise is any kind of increased uncertainty about the intended signal, not just acoustic noise. Other work suggests that talkers provide redundancy that guards against non-acoustic noise, such as perceptual confusability and plausible minimal pairs. For example, Zhao and Jurafsky (2009) found that speakers of Cantonese, a language with six lexical tones, produced elevated f0 for all tones in the presence of noise (in an expected Lombard effect), but found an effect of word frequency on f0 only for mid-level tones. Zhao and Jurafsky pointed out that these are the tones that are most perceptually confusable (Khouw and Ciocca, 2007; Vance, 1977; Whitehill et al., 2000), and so benefit the most from additional redundancy. Additionally, as mentioned in the introduction, Buz et al. (2014) found that talkers produce stronger VOT contrasts when there is a plausible VOT-minimal-pair referent in the discourse context, a result which Kirov and Wilson (2012) also found with a different experimental methodology. All together, these effects suggest that talkers will provide redundancy that is especially suited to guarding against a communication error when the channel degrades.

There is also some evidence that talkers take advantage of *increased* channel capacity. In “Map Task” studies, a “guide” tells a “follower” how to draw a path on a map, and some landmarks on the guide’s map are missing from the listener’s map (Bard et al., 2000). In Boyle et al.’s 1994 original Map Task dataset, some pairs of talkers could see each other, and others could not. Boyle et al. found that talkers with the additional visual channel completed the task more quickly and at the same level of accuracy than talkers without that channel, communicating the same amount of information in less time. Moreover, Boyle et al. found that pairs in the Visible condition sought to establish eye contact more often during periods of communicative difficulty, suggesting that visibility increases channel capacity in a way that talkers seek to exploit.

It is worth noting that many predictability effects cannot be attributed unambiguously to either source or channel coding, but could plausibly be the result of either, or both together. This ambiguity of interpretation is due to a fundamental confound that arises in language behavior: redundancy is always present to some extent in linguistic productions, and the elements that should receive longer signals under source coding should also receive more redundancy under channel coding. To eliminate redundancy in source coding, we should seek to pair probable meanings with short signals. At the same time, to avoid errors, we should strategically insert redundancy to, or preserve pre-existing redundancy to, those parts of the signal that are most likely to give rise to errors (channel coding). Since listeners likely have more trouble processing low-probability items, this means that channel coding should also make low-probability items longer and more redundant. Thus, simply finding an inverse correlation between the probability of some linguistic element and its length or level of redundancy, without determining whether the redundancy is especially effective at avoiding errors, is not enough to determine whether the correlation reflects a tendency towards source coding, channel coding, or both.

2.3. Predictability effects according to particular listener characteristics

As mentioned, an ideal talker could make speech more efficient by using listener characteristics (i.e., their knowledge of what the listener knows or is like) to inform both their assumptions about how probable different messages are (for better source coding), and their estimates of how probable different errors are (for better channel coding). The idea that talkers tailor their utterances to be understood by their audience, called *Audience Design*, has its roots in Grice's maxims that talkers endeavor to be unambiguous but not overly informative (Grice, 1975). Clark and collaborators further developed this idea by proposing that talkers maintain a model of the *common ground*, which consists of information that has been shared with all involved interlocutors, and consult this common ground in designing utterances that can be understood easily (Clark and Carlson, 1982; Clark, 1996; Clark and Marshall, 1981). The common ground represents knowledge of a particular set of listener characteristics that talkers might use to modulate their productions—specific facts the listener knows or perceives.

Brown and Dell (1987) and Dell and Brown (1991) elaborated on this idea to distinguish between *generic* adjustments that adjust for an average listener, and *particular* adjustments that dynamically adjust to the details of the current listener. Dell and Brown observed that generic adjustments could easily arise if talkers use the same systems for accessing linguistic items for both production and comprehension: talker demands will be similar to average listener demands, and items that are easy to understand will be the same as those that are easy to produce. This possibility provides a plausible route for generic listener effects to arise not in response to adaptive pressure but as a 'spandrel' side-effect of the same systems being involved in production and comprehension. Particular listener effects, on the other hand, should not arise under this account.

Brown and Dell (1987) performed an experiment in which subjects read stories silently, and re-told the stories to a confederate. Some stories involved typical instruments, such as 'stabbing' with a 'knife,' but others involved unusual instruments, such as 'stabbing' with an 'ice pick.' The stories were accompanied by illustrations that included the instrument. In the crucial manipulation, in one condition, both the subject talker and confederate listener could see the illustrations, but in the other condition only the talker could see the illustrations. Brown and Dell (1987) found that talkers mentioned atypical instruments more often than typical ones, but ignored the availability of the illustration to the listeners. This result suggested that talkers responded to their own view of the typicality of the instrument, but did not care if the listener had access to the typicality of the instrument. Lockridge and Brennan (2002) observed that confederate listeners who have heard the same stories many times may appear unengaged to talkers, and repeated the experiment with naïve listeners. With this approach, contrary to Brown and Dell (1987), they found that talkers did provide more information about atypical instruments when listeners could not see the illustrations. When listeners could not see the illustration, talkers were more likely to mention atypical instruments at all, to mention them earlier, and to use indefinite references. By providing evidence for particular listener adaptation in addition to generic listener adaptation, this result supports the view that talker behaviors that appear to support communicative efficiency reflect an overall pressure towards efficient communication. Also see Roche et al. (2010), who found that talkers avoided ambiguity only when they believed a pre-scripted pseudo-confederate to be a real person.

Many subsequent studies on audience design have focused on the case of referring expressions, where results suggest that talkers sometimes seem to account for listener knowledge in determining the form of these expressions, but not in all cases. For example, some studies found evidence for a more restricted notion of audience design that relied on an "ego-centric" model of predictability. In an image description task, Fukumura and van Gompel (2012) found that talkers tend to use pronouns, rather than definite noun phrases, when a prompt mentioned the target, even if the listener did not hear the prompt (because the talker heard it through headphones). In Bard et al.'s 2000 Map Task study mentioned above, they found that talkers tended to refer to any given landmark on the map with shorter referring expressions after the landmark had already been mentioned (and so was more probable). However, the extent of the guide's reduction did not change depending on who provided the initial mention, even though a guide-mentioned landmark may not be present on the listener's map.

Other studies have found evidence for a more expansive notion of audience design, in which predictability is assessed from the listener's perspective For example, talkers are more likely to include a disambiguating adjective ("the large circle") when a competitor referent (small circle) is visible to the audience (Horton and Keysar, 1996), and are more likely to use a full noun, rather than a pronoun or null word, when the audience cannot see the referents (Mathews et al., 2006). When talkers know a (short) name for a referent, they tend to use the name if their listener also knows the name, but use a (long) description if the listener does not know the name (Heller et al., 2012). In addition, when talkers are asked to tell the same story twice, either to the same listener or to different listeners, their second telling includes more narrative elements, and more words per element, if the listener has not heard the story before (Galati and Brennan, 2010).

Overall, these results suggest that talkers can consider particular listener characteristics in determining the probability of different elements, but that the computational difficulty of tracking very detailed characteristics may interfere. Galati and Brennan (2010) proposed that their talkers exhibited audience design because they need only a “one-bit” model that indicates whether the audience has heard the overall story, and Fukumura and van Gompel (2012) proposed that their talkers did not exhibit audience design because doing so would involve tracking the accessibility of each individual referent. Other studies have indicated that reducing the cognitive load of a task (by giving participants more practice or allowing more time) can result in stronger effects of audience design (Horton and Gerrig, 2002, 2005; Roche et al., 2010), supporting the idea that audience design may be computationally expensive. In a refinement of this idea, Bard and Aylett (2005) proposed the “Dual-Process” model in which listener characteristics can be considered in the utterance-planning stage, but are too computationally expensive to consider during low-level articulation. Their proposal was based on a reanalysis of the Map Task data (Bard et al., 2000) showing that speed of pronunciation did not change according to the listener’s knowledge, but that the form of the referring expression (e.g., pronoun or full noun phrase) did. Alternatively, we can turn to Jaeger and Ferreira’s (2013) proposal that efficient communication involves context-specific learning. Under this account, decreased cognitive load eases audience design because context-specific learning is easier, not because talkers manage to engage a discrete process. Roche et al. (2010), in an object arranging task, and Roche et al. (2014) , in a collaborative toy construction task, found that participants modified their referring expressions only after salient communication errors, providing evidence that at least some audience design is a consequence of context-specific error-driven learning.

Together, these results indicate that talkers are capable of ‘particular listener’ audience design, but exhibit audience design only if it is likely to have an impact (Jaeger and Ferreira, 2013; Jaeger, 2013; Kurumada and Jaeger, To Appear). In some cases, assessing listener needs is either so easy or so hard that the choice is clear. When it is easy to assess listener characteristics, as in the ‘one-bit’ case, talkers clearly commit to adjusting to listener needs; when it is hard, requiring the talker to track the accessibility of a set of referents to a listener, talkers do not commit. In cases where the decision is not so clear, talkers gather more information to evaluate this trade-off. For example, talkers will visually assess the level of the listener’s engagement, and adjust to engaged naïve listeners but not to disengaged confederates.

Nevertheless, for particular-listener characteristics that are easier to track, the results of these experiments accord with the predictions of an information-theoretic account. On the other hand, there is so far little, if any, evidence that talkers account for particular-listener characteristics in predictability effects at the phonetic level. Our Analysis I looks at a listener characteristic that is extremely easy computationally: whether the listener is an infant or an adult. This analysis arguably considers a third kind of listener adaptation. Infants are certainly not the 'average listener' that a generic adjustment would target, but it is not clear that they present the kind of specific information imbalance that the previous studies have considered. Instead, these infants are overall not competent speakers. Regardless, we find evidence that this listener characteristic affects not only overall phonetic realization (e.g., speech rate), but differentially affects high-probability and low-probability words—that is, the effects of predictability differ depending on the listener, even on low-level articulatory processes.

Having found an effect of listener characteristics on predictability effects, we will look for an effect of channel characteristics. As discussed above, previous work on the Lombard effect shows that talkers guard against degraded channels (whether consciously or not): when talkers hear noise, they produce louder, more intelligible speech. Our Analysis II will consider whether talkers can also take advantage of enhanced channels. Using the same Map Task dataset mentioned above, we will show that the talkers who could see each other achieved their higher information rate in part by providing less acoustic redundancy. Together, these results provide evidence that predictability effects reflect adaptation to efficient communication, and are not simply an interesting, but ultimately unadaptive, side-effect of how neural hardware works.

3. Analysis I – infant- and adult-directed speech.

In this analysis, we will compare predictability effects in infant-directed speech (IDS) and adult-directed speech (ADS) to see if coarse listener characteristics influence predictability effects at the phonetic level. Our general methodology is borrowed from the corpus study of Bell et al. (2009) investigating predictability effects. The current study differs from Bell et al. (2009) in comparing predictability effects in ADS and IDS, using mixed effects regression rather than standard regression, and residualizing against a control model to minimize collinearity between predictors of interest and control predictors.

3.1. Data

We use data from two corpora: `swbdnxt` (Calhoun et al., 2010), an edition of the Switchboard corpus of telephone conversations between adults, and `Large Brent` (Rytting et al., 2010), a subset of the Brent

Table 1: Statistics for the two datasets used in Analysis I.

	# Sent	# Words	$\frac{\text{Word}}{\text{Sent.}}$	# Talkers
swbdnxt	1,985	18,621	9.4	75
brent	2,254	14,148	6.3	4

corpus (Brent and Siskind, 2001) of spontaneous infant-directed speech. We describe these corpora and the data extracted from them below.

3.1.1. swbdnxt

swbdnxt is an edition of the Switchboard corpus of telephone dialogues between adults. It integrates several levels of annotation produced by different groups since the original Switchboard release. These include prosodic and syntactic annotations, as well as a phonetic alignment created by correcting the output of a forced alignment produced using a pronunciation dictionary. The prosodic annotations follow the Tones and Break Indices (ToBI) transcription standard (Beckman et al., 2005); we use only the break indices in this work.

To create the Switchboard dataset for our analysis, we used the Mississippi State transcript of the Switchboard data, and began with all words that were annotated with Penn Treebank part-of-speech (POS) tags and Mississippi State ToBI labels. The resulting corpus had $69,887$ words in $6,593$ utterances across 75 conversation. We used this dataset to compute the Adult-Directed version of the language model, as described shortly.

We further processed the Switchboard data to produce a dataset that was as comparable to the Infant-Directed corpus as possible. To avoid having to handle correlations between talkers in the same conversation, we discarded all sentences from side B, producing a dataset with $36,273$ words in $3,420$ utterances. To avoid disfluencies, we discarded all words that were annotated with a POS tag of “UH” or “XX”, eliminating $3,006$ words and 820 utterances. We then removed all words that were annotated as part of a ‘disfluent’ prosodic phrase, discarding a further $6,066$ words and 108 utterances. Because formulaic backchannel responses may have very different durational properties, we also discarded all words that were annotated as part of a ‘backchannel’ prosodic phrase, removing 701 words and 32 utterances. Finally, following Bell et al. (2009), who note that short prosodic phrases are formulaic discourse responses, we discarded all prosodic phrases that were shorter than 4 words. Table 1 shows statistics for the final swbdnxt corpus.

3.1.2. `Large Brent`

`Large Brent` is a subset of the Brent Corpus of spontaneous IDS collected in a naturalistic setting. It consists of the mothers' utterances from four mother-infant dyads, and has a forced phone alignment based on a modified version of the CMU pronunciation dictionary. Details of the corpus and alignment can be found in Rytting et al. (2010). `Large Brent` has a 90%/10% train/test partition; for this study we use only the training partition, which contains $22,226$ words from $7,030$ sentences. Rytting et al. have already excluded utterances containing partial or unintelligible words, so we made no further effort to handle disfluencies.

Unlike `swbdnxt`, this corpus does not include talker's ages; since all talkers are new mothers, we use an estimate (based on personal communication with Michael Brent) of 27 years old for all talkers. There is also no annotation of intonational phrase boundaries, which are known to affect word duration. However, in this corpus every pause of 300 milliseconds or more is taken to be an utterance boundary, so we use the utterance boundaries as a fairly robust approximation to intonational phrase boundaries. As in the ADS corpus, we remove all prosodic phrases which are three words or shorter, resulting in the corpus statistics shown in Table 1.

3.1.3. Pooled dataset

To facilitate direct comparison between predictability effects in ADS and IDS, we created a pooled dataset containing the data from both `swbdnxt` and `Large Brent`. As can be seen in Table 1, the pooled dataset is relatively balanced in terms of the number of sentences and words from each type of speech, but not in terms of talkers. The imbalance in the number of talkers is handled by including random effects for Talker in our model when the random effect is a significantly better fit (described below).

3.2. Models

All of our models are mixed-effects models. For general introductions to mixed effects models, we refer readers to Baayen et al. (2008), and for an introduction to using mixed effects models with corpora, see Bresnan et al. (2007) or Jaeger (2010). For the more specific case of using mixed effects models with phonetic corpora, see Tily et al. (2009) or Kuperman and Bresnan (2012).

3.2.1. Approach

Word duration is affected by many factors other than word predictability, such as talker age, speech rate, the word's length in phones, and its position in the intonational phrase. To control for these kinds of factors, we adopt a simple two-step regression procedure, illustrated in Figure 1. First, we build a single control

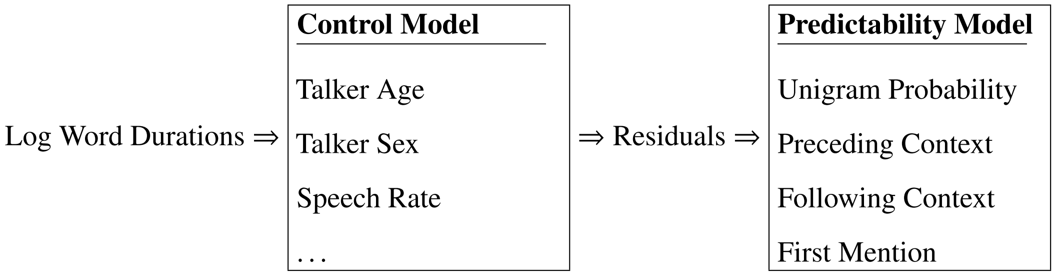

Figure 1: Schematic of two-step modeling process.

model (using the model selection procedure described shortly), which regresses log word duration against only control terms such as those above.[3] The control model is fitted to the pooled dataset, and includes a Speech Type term (ADS or IDS) to allow for effects on duration that are due to speech type but not related to predictability. We also allow interactions between Speech Type and the other control factors. Next, we take the residual log durations of this control model as the response variable for model selection among predictability terms, so that these terms can only be used to explain the part of the variance that has not already been accounted for by control factors. We perform three separate regressions on the residuals: one on the residuals from the ADS subset of the data, one on the IDS subset, and one on the entire pooled dataset. The ADS and IDS regressions allow us to assess which predictability terms are significant factors in predicting word duration in ADS or IDS, while the pooled regression can show, through interactions with the Speech Type term, whether a given predictability factor has different effects in ADS and IDS.

This two-step approach has the advantage that we do not need to worry about collinearity among our control predictors. If two predictors are collinear, the parameters of the terms in the control model will be unstable, but the overall variation explained, and the residuals, will be stable.

3.2.2. Fixed Effects: Control terms

We include nine terms in the control model, based on those of Bell et al. (2009). These are: **Talker Age** (taken from metadata in `swbdnxt`; estimated as described in the Data section for `Large Brent`), **Talker Sex**, **Speech Rate** (computed per utterance as $\log\left(\frac{\text{\# Vowels}}{\text{Second}}\right)$), **# Vowels** (taken from annotation based on a pronunciation dictionary), **Log Average Expected Word Duration** (a log-space uniphone model, computed as the log of the sum of the average duration of each phone in the word, following Bell et al. (2009); average

[3] Like Bell et al. (2009), we take the log of the duration to avoid equating a 50ms difference in a word that is usually 60ms with a 50ms difference in a word that is usually 300ms.

phone durations were computed separately for each dataset), **Intonational Phrase Initial** (indicates whether a word is at the beginning of an intonational phrase; phrases are bounded by TOBI break indices of 4 in `swbdnxt` and are assumed to coincide with utterances in `Large Brent`), **Intonational Phrase Final** (as previous), **Content or Function Word**[4] (based on POS tags), and **Speech Type** (ADS or IDS).

3.2.3. Fixed Effects: Predictability terms

We include four predictability terms, again following Bell et al. (2009): **Log Unigram Probability**, **Preceding Context** (log probability of a word given the preceding two words), **Following Context** (log probability of a word given the following two words), and **First Mention** (whether or not the word has appeared in the transcript). While First Mention can be read off the transcripts, the first three predictors are the output of language models that must be estimated from data. Because these measures should reflect the predictability of these words for these talkers, we estimated the language models using transcripts from the same corpora used in our analyses. We built one set of language models for ADS using the `swbdnxt` data (including both sides A and B as noted in Section 3.1.1; the additional data will lead to better language model estimates), and one set for IDS using a superset of the `Large Brent` data. Specifically, we used six mother-child dyads (c1, f1, f2, i1, j1, and q1) from the Brent Corpus, which collectively contain roughly the same number of words as were available for the ADS language model (75,184 words in 24,272 utterances across 40 recording sessions).

We will be performing two kinds of regressions; an individual regression for each speech type, and a pooled regression. For our individual regressions on one speech type only, we used Good-Turing smoothed language models that were estimated from the corresponding corpus. For our pooled regression, we need each predictability term to be based on a single language model, i.e., the Unigram Probability of a particular word will not depend on whether it is from the IDS or ADS corpus. However, to avoid biasing the pooled model, we also need each language model (the Unigram Probability model, the Following Context model, and the Preceding Context model) to fit each corpus equally well.[5] That is, the average probability of the

[4] Bell et al. (2009) investigate this term in interaction with predictability terms. We attempted to include it in our predictability model, but it is highly collinear with other predictability terms and we failed to reduce collinearity to an interpretable level, so we include it in the control model instead.

[5] See Fine et al. (2014) for evidence that language model bias can affect how well a language model predicts behavioral responses. In practice, we found the same pattern of significant effects regardless of whether we used the procedure explained below to fit λ interpolation parameters, or just weighted each model equally ($\lambda = .5$) in the pooled model.

words in the IDS corpus (as estimated by the model) should be equal to the average probability of the words in the ADS corpus. To achieve both of these goals at once, we created a single language model for each predictability term by taking a weighted average of the IDS and ADS models. For example, for Preceding Context, we computed:

$$P_{pooled}(w_i|w_{i-1}, w_{i-2}) = \lambda P_{IDS}(w_i|w_{i-1}, w_{i-2}) + (1 - \lambda)P_{ADS}(w_i|w_{i-1}, w_{i-2}) \tag{2}$$

where λ is a scalar between zero and one that is chosen to provide equal perplexities (a measure of model fit) on the ADS and IDS data sets. For Unigram Probability, $\lambda = 0.250$; for Preceding Context, $\lambda = 0.661$; for Following Context, $\lambda = 0.667$. Concretely, the ADS model is weighted more strongly in the pooled Unigram model, while the IDS model is weighted more heavily in the pooled context models.

3.3. Model Selection

We employ a forward model selection procedure that closely follows an algorithm introduced by Coco and Keller (2010), with only minor modifications to avoid specific interactions that lead to convergence errors (all our IDS talkers are female, for example, so we avoid testing for an interaction between Speech Type and Sex).[6] For each round of model selection, we consider two random effects: Talker, and Sentence. We first determine which of the two random effects (intercept only) produces a better initial model. We then determine whether adding the other random effect (intercept only) produces a significant improvement in model fit. This is followed by another series of model comparisons to add fixed main effects and random slopes for each random effect (including random effects that are not already in the model). Finally, we perform another series of model comparisons to add interactions.

In each step, a predictor is added if the model with that predictor is a significantly better fit to the data than the model without that predictor, as assessed with the `anova` function for model comparisons in `R`. In the results to be presented, we explored models that assumed random slopes and intercepts were conditionally independent, given the random variable.

We are performing model selection only because a model that expresses the full range of interactions and random effects does not converge. Our goal is then to obtain the largest (and so least anticonservative, Harrell, 2001, Chapter 4) model that converges, not to find some "true" model. Accordingly, during model selection we are eager to add terms, and set a relatively loose criterion for adding terms. If a larger model is

[6] The R implementation of the modified algorithm is available at http://jkpate.net/modelselect.R.

a better fit with $p < 0.1$, we adopt the larger model. All predictors were centered except Speech Type; for ease of interpretation, Speech Type was set to -1 for adult-directed Speech and 1 for infant-directed speech, resulting in a mean value of ≈ -0.137.

To obtain p-values, we do one final round of model comparisons. For each fixed effect, we fit a model that contains all selected predictors except that fixed effect, and compare that model with the full model. Our presented tables contain the χ^2 values and the corresponding p-values from this model comparison.

3.4. Results

3.4.1. ADS and IDS individually

Before performing a direct comparison of ADS and IDS data, we first build individual predictability models for the two types of speech. The individual models serve two purposes. First, they tell us what kinds of effects are evidently present in each kind of speech, allowing an informal comparison of the predictability effects in ADS and IDS speech. Second, we can use the results of these individual models to inform model selection when performing a direct comparison on the pooled data. In short, the individual models identify patterns of significant effects, and the pooled model compares these patterns of significant effects in a quantitative manner.

For the individual models, we take the residuals from the control model fitted on the entire pooled dataset, and run model search on the IDS predictability terms to predict the residuals for IDS words, and then run model search on the ADS predictability terms to predict the residuals for ADS words. We restrict model search to main effects.

Results for ADS and IDS are presented in Tables 2a and 2b, respectively. As the response variable is (residual) log duration, a negative coefficient corresponds to a shortening effect as the predictor increases, and a positive coefficient corresponds to a lengthening effect. For both ADS and IDS, a number of predictability effects are observed. For ADS, we find most of the expected predictability effects among the main effects: frequent words and words predictable from context are shorter. However, there is no apparent effect of First Mention.[7] In IDS, we also find shortening effects of Unigram Probability and Following Context, but

[7] This conflicts with some previous work. Bard et al. (2000) and Bard and Aylett (2005) found that words were pronounced with shorter duration the second time they appeared in the conversation, and Aylett and Turk (2004) and Bell et al. (2009) found an overall shortening effect of repetition count and log repetition count, respectively. However, these studies did not use mixed effects models, and considered a narrower range of control predictors. Gahl et al. (2012) performed a mixed effects analysis on the Buckeye corpus (Pitt et al., 2007), and, similar to our results, did not find an effect of First Mention on word duration or vowel dispersion.

Table 2: Fixed effects for individual ADS and IDS regressions, with χ^2 statistics and p-values from model comparisons.

Predictability Term	Coeff.	χ^2	p-val
(Intercept)	0.0097	15.9	0.0001
Unigram Prob	-0.0225	58.9	0.0001
Prec. Context	-0.0056	27.1	0.0001
Foll. Context	-0.0062	37.8	0.0001
First Mention	N/A	N/A	N/A

(a) Coefficients of the individual ADS model.

Predictability Term	Coeff.	χ^2	p-val
(Intercept)	0.0197	18.1	0.0001
Unigram Prob.	-0.0276	141.2	0.0001
Prec. Context	N/A	N/A	N/A
Foll. Context	-0.0069	13.3	0.0003
First Mention	N/A	N/A	N/A

(b) Coefficients of the individual IDS model.

Preceding Context was not added to the model (and it is not significant if its addition is forced). We discuss the implications of this difference in Section 3.5.

These individual models, however, only reveal whether we have enough evidence to determine that the various coefficients are significantly different from zero, without comparing the effects in ADS with those in IDS. A direct comparison accomplishes two goals. First, it can confirm that the effect of Preceding Context is actually weaker in IDS than it is in ADS. Second, it is possible that an effect might be significant and in the same direction in both types of speech, but be much stronger in one type of speech. A pooled model on the entire dataset enables just such a direct comparison of effects in each Speech Type by examining interactions with the Speech Type term. We now proceed to this pooled model.

3.4.2. Pooled comparison

In this section, we perform model search on the full pooled dataset, and include in the predictability model the fixed effect "Speech Type" that indicates whether each word is from the ADS dataset or the IDS dataset. We will in particular be examining the interaction terms between Speech Type and the predictability terms. Since we wish to verify different patterns of significant results in models containing only main effects, we consider only interactions involving Speech Type. Thus, we fix in advance the main effects, and perform model search only for interactions and random effects.

Table 3: Fixed effects for pooled regression, Speech Type coded with IDS as 1, with χ^2 statistics and p-values from model comparisons.

	Predictability Term	Coeff.	χ^2	p-val
Main effects	(Intercept)	0.0021	0.7	0.4022
	Unigram Prob.	-0.0233	69.1	0.0001
	Prec. Context	-0.0027	4.6	0.0320
	Foll. Context	-0.0065	29.4	0.0001
	First Mention	0.0050	2.9	0.0857
	Speech Type	-0.0001	0.1	0.9585
Interactions	Speech Type × Prec. Context	0.0041	10.8	0.0010
	Speech Type × Unigram Prob.	N/A	N/A	N/A
	Speech Type × Foll. Context	N/A	N/A	N/A
	Speech Type × First Mention	N/A	N/A	N/A

Table 3 presents the model coefficients and p-values from our final model. The table is separated into one box for main effects, and another box for interactions between predictability terms and speech type. As our Speech Type variable is approximately centered (IDS is 1 and ADS is -1), the main effects terms indicate an approximate average over the entire pooled corpus (with a slight bias to ADS). We observe in the main effects a shortening effect of Unigram Probability and both Preceding and Following Context. The coefficient for First Mention, which was not significant in either individual regression, is positive, suggesting a trend towards a lengthening effect for new items. However, the term again does not reach significance.

There is only one significant interaction with Speech Type; we see a significant and positive interaction between Speech Type and Preceding Context, indicating that the shortening effect of Preceding Context is stronger in ADS.

3.5. Discussion

These results show some similarities in the effects of predictability on word duration in IDS and ADS. Both speech types exhibit significant shortening according to Unigram Probability and Following Context. However, our results also revealed a difference: while there is a shortening effect of Preceding Context in ADS, there is not an apparent effect of Preceding Context in IDS.

Before discussing this specific difference between ADS and IDS, we first comment on the fact that IDS exhibits predictability effects at all. At first glance, this result might seem to support the idea that predictability effects are *not* about efficient communication of linguistic messages, since it is not obvious that efficiency in communication is a goal of parent-infant interactions—speech directed to pre-verbal infants may be more about social bonding or communicating the nature of the grammar, for example. On the other hand, predictability effects in IDS are not surprising in a 'spandrel' account where they are a side effect of general cognitive mechanisms but carry little communicative benefit. Under such a theory, the same mechanisms that cause predictability effects in ADS would also do so in IDS. Of course, this theory also predicts that we should see the *same* predictability effects in ADS and IDS, which we did not. The difference we found does not fall out simply because IDS is slower than ADS; we found a difference in the effect of Preceding Context on word duration between the speech types, not merely an overall difference between the speech types.

To account for this difference according to listener, we would need more spandrel accounts. For example, we may have found an effect of Preceding Context only in the ADS data because the ADS conversations were more cognitively demanding overall and occupied enough resources to reveal an effect of production ease, while the abundance of fungible computational resources for the IDS talkers (perhaps due to the slower speech rate in IDS) led to a floor effect in production difficulty. In any event, the point is that while merely observing predictability effects in IDS is not a problem for a spandrel account, the particular pattern of results we obtained apparently demands multiple independent spandrel accounts.

Turning to the information-theoretic account, there are two possible explanations for why we might observe predictability effects in IDS. It could be that efficient communication really is the goal of parent-infant interactions, so many of the same effects we see in ADS will also appear in IDS (but potentially with some differences due to the different linguistic capabilities of the listeners). Alternatively, even if efficient communication is not the goal of IDS, it could be that our language production systems are optimized in a "listener-general" way that is efficient for standard communicative scenarios (i.e., adult conversation), but have difficulty adapting dynamically towards the efficiency needs (or lack thereof) of other "listener-specific" scenarios (Dell and Brown, 1991). Note that the latter hypothesis is very difficult to distinguish empirically from the view that predictability effects are 'spandrels': whether they reflect generic adaptation or no adaptation, we would expect to see no modulation of predictability effects in response to differences in the listener or the channel. However, since we did find differences between ADS and IDS, predictability effects apparently *do* vary according to the listener, at least to some extent. This result extends previous findings that overall speech style varies for different listeners, such as foreigners (Uther et al., 2007; Papoušek and Hwang, 1991) and infants (Fernald and Simon, 1984; Werker et al., 1994; Martin et al., 2014), to suggest that the details of how predictability influences word duration varies for different listeners. Thus, our findings cast doubt not only on the spandrel theory but also on the theory of adaptation towards average scenarios with no dynamic adaptation.

Regardless of whether IDS is subject to similar efficiency pressures as ADS, we are left with the task of explaining why we obtained the particular pattern of differences that we did. Fully answering this question will require considerable further research (which is also important to confirm the particular pattern of results we found), but we provide a speculation here. The overall pattern we observe in the IDS is a diminished effect of preceding context. This pattern may reflect different communicative goals for ADS and IDS: ADS is focused on structures that involve several words, but such structures are less important in IDS. These results suggest that talkers can, subconsciously, "turn off" at least some predictability effects, and the particular pattern suggests that they may selectively turn off predictability effects that are not relevant to the demands of the current situation.

Whether or not this explanation is correct, the results of Analysis I suggest that predictability effects change for different listeners, bolstering the view that predictability effects reflect adaptation to communication. We will next consider how predictability effects change in the presence of a visual channel, which presumably increases channel capacity. This second study also serves to control for a potential confound in Analysis I, since the IDS talkers could see their listeners but the ADS talkers could not. In Analysis II, we will examine the effect of listener visibility on predictability effects, comparing dialogues that were recorded in identical conditions, except some talkers could see their listener, and some talkers could not.

4. Analysis II – The effect of speaker visibility

Under an information-theoretic efficiency account, communication with a visible partner presumably increases the capacity of the communicative channel. For example, when giving directions, a talker may use hand gestures to reinforce the approximate angle of a particular turn as "veering" rather than simply turning. Remember from the Background section that the capacity of a channel is the entropy of the source coding distribution $H(X)$, minus the entropy of the channel-coding probability distribution over signals after noise has been applied $H(X|Y)$:

$$H(X) - H(X|Y)$$

The second quantity, $H(X|Y)$, reflects the average amount of reinforcement of the intended signal that is needed to resist noise.

In the general case, the presence of a visual channel may change both $H(X)$ (people who can see each other may tend to talk about different things) and $H(X|Y)$ (people who can see each other are subject to different noise). If, for the purposes of argument, we hold $H(X)$ fixed, however, then talkers with a visible partner will enjoy an equal or increased channel capacity compared to talkers with a non-visible partner, for the following reason. In the worst case, the visible cues, such as gestures and facial expressions, will be completely uncorrelated with the linguistic signal, in which case $H(X|Y)$ is the same in both situations, and the channel capacities are equal. To the extent that visual cues do correlate with the linguistic signal, however, $H(X|Y)$ will be smaller (less ambiguous) with a visible partner, increasing the difference above and so the channel capacity.

There are already reasons to think that gestures correlate with the linguistic signal. Cook et al. (2009) elicited productions by asking participants to describe short videos, and recorded whether participants produced a gesture when producing a dative verb. When talkers produced a verb in the verb's dispreferred form of the dative alternation, they were more likely to produce a hand gesture. Esteve-Gibert and Prieto (2013) examined the time course of hand gestures during a point-and-name task in Catalán, and found that the apex of the pointing gesture tended to align with the fundamental frequency peaks of stressed syllables, although there was a complicated interaction with foot structure. Similarly, Graf et al. (2002) found that talkers align the extrema of their head tilts with fundamental frequency peaks of English pitch accents. These results suggest that talkers can use visual cues to provide redundancy, and so there is an opportunity for talkers to offload redundancy from the acoustic channel to the verbal channel.

In the Map Task dataset we will use, described shortly, the messages will be controlled so that $H(X)$ is the same with and without a visual channel: the talkers are ultimately trying to communicate the same paths on the same maps. Previous work has found that these Map Task talkers communicated more quickly with no decrease in accuracy when they have a visible partner. We will replicate this finding, and also find evidence

Table 4: Statistics for the Map Task dataset used in Experiment II.

	# Sent	# Words	$\frac{\text{Word}}{\text{Sent.}}$	# Talkers
maptask	7,592	77,485	10.2	64

that some predictability measures that correlate with phonetic redundancy in the non-visible condition show no such correlation in the visible condition. This result indicates that the talkers rely mostly or entirely on the visual channel for error avoidance redundancy when it is available, but will adapt their pronunciation to provide the needed redundancy when a visual channel is not available.

4.1. Data

For this analysis, we use data from the same HCRC maptask corpus used by Boyle et al. (1994) and Bard et al. (2000). This dataset is a collection of transcribed recordings of unscripted, task-oriented dialogue. Each dialogue has two participants, and they are both given a cartoon map with several labeled landmarks. One participant, the "guide," is given a map with a route drawn on it, while the other participant, the "follower," is given a map with no route. The guide and follower then converse to help the follower draw the route on his or her own map. In half of the dialogues, the guide and the follower can see each other, while in the other half they cannot. In neither condition did they see each other's maps. This dataset contains hand-annotated start and end times for each word.

To prepare the corpus for our study, we started with all dialogues, forming an initial corpus with $20,719$ sentences ($145,483$ words). We then discarded sentences with fewer than four words, producing a dataset with $12,025$ sentences ($132,076$ words). Next, we discarded any word that had been annotated as part of a disfluency, producing a dataset with $11,782$ sentences ($107,759$ words). Finally, we discarded all follower utterances, leading to the dataset statistics in Table 4.

4.2. Models

We used the same basic modeling approach as in Analysis I, using the same model selection procedure to first fit a control model, and then performing model selection to find a regression of the residual log durations from the control model against the predictability factors. We estimated the language models in the same fashion, using Good-Turing smoothing for the individual Visible and Non-visible regressions, and interpolating language models to control for model fit for the pooled regression. For Unigram Probability,

λ = 0.629 (the Visible model was weighted more heavily); for Preceding Context, λ = 0.489; and for Following Context, λ = 0.484. The set of control predictors for Analysis II were essentially the same as the control predictors for Analysis I. The corpus annotation provides the age of each talker, so we used the real age of the talker. Only a small part of the `maptask` corpus has ToBI-style break index labels, so, as with the `Large Brent` corpus, we assumed that utterance-final (-initial) words were the only prosodic phrase-final (-initial) words. Because of the structure of the Map Task corpus and available metadata, Analysis II involved more random effects. The model selection function considered random intercepts and slopes according to sentence, talker, conversation, map, and talker birthplace.

4.3. Results

As in Analysis I, we first build individual predictability models for the Visible condition and the Non-Visible condition. These individual models will provide an intuition for which effects are and are not present in each Visibility condition. Similarly, to verify whether differences in patterns of significant effects are real, we will fit a full model on both conditions, and see if interactions with Visibility are significant.

4.3.1. Visible and Non-Visible Individually

Results for the Non-Visible and Visible conditions are presented in Tables 5a and 5b, respectively. As before, the response variable is (residual) log duration, and a negative coefficient corresponds to a shortening effect as the predictor increases. Among the Non-Visible results of Table 5a, we see the expected shortening effects of Unigram Probability, Preceding Context, and Following Context. Additionally, we see a lengthening effect of First Mention, but it is not significant. Looking to the Visible results of Table 5b, we see what appears to be a similar pattern, with shortening effects of Unigram Probability, Preceding Context, and Following context, but First Mention was not added to the model.

These regressions suggest that the same effects exist, regardless of the presence of a visual channel. However, it is still possible for there to be a difference in magnitude due to the presence of a visual channel. We turn next to a pooled comparison to test for such a difference.

4.3.2. Pooled comparison

Similarly to Analysis I, we fit the pooled model on the concatenation of the Visible and Non-Visible datasets, and set the fixed effects to be the predictability measures, Visibility, and interactions between the fixed effects and Visibility, performing model search only among interactions and random effects.

Table 5: Fixed effects for individual Visible and Non-Visible regressions, with χ^2 statistics and p-values from model comparisons.

Predictability Term	Coeff.	χ^2	p-val
(Intercept)	0.0002	0.1	0.9446
Unigram Prob.	-0.0333	116.1	0.0001
Prec. Context	-0.0109	89.1	0.0001
Foll. Context	-0.0277	74.4	0.0001
First Mention	0.0053	2.9	0.0908

(a) Coefficients of the Invisible model.

Predictability Term	Coeff.	χ^2	p-val
(Intercept)	-0.0006	0.1	0.7770
Unigram Prob.	-0.0261	61.0	0.0001
Prec. Context	-0.0133	37.5	0.0001
Foll. Context	-0.0257	128.0	0.0001
First Mention	N/A	N/A	N/A

(b) Coefficients of the Visible model.

Table 6 presents coefficients and p-values from our final model. Among the main effects, representing an average across the Visible and Non-Visible data, we see significant shortening effects of Unigram Probability, Preceding and Following Context, but no effect of First Mention.

We see one interaction with Visibility, and it is a significant ($p \leq 0.05$) positive interaction between Visibility and Unigram Probability, indicating that any shortening effect of Unigram Probability is significantly weaker in the Visible condition, and stronger in the Non-Visible condition. Thus, while all three predictability measures have an effect, regardless of the presence or absence of a visual channel, Unigram Probability has a weaker shortening effect in the presence of such a channel.

4.4. Discussion

This result reveals a difference in the magnitude of reduction between the Visible and Non-Visible cases; while probable words are shorter in both cases, Unigram Probability has a smaller effect in the presence of a visual channel. This result first indicates that the difference we found between ADS and IDS in Analysis I

Table 6: Fixed effects, visibility coded with Visible as 1, with χ^2 statistics and p-values from model comparisons.

	Predictability Term	Coeff.	χ^2	p-val
Main effects	(Intercept)	0.0002	0.1	0.9176
	Unigram Prob.	-0.0280	197.5	0.0001
	Prec. Context	-0.0148	77.3	0.0001
	Foll. Context	-0.0288	240.7	0.0001
	First Mention	0.0065	4.8	0.0280
	Visibility	-0.0001	0.1	0.9570
Interactions	Visibility × Unigram Prob.	0.0027	4.6	0.0318
	Visibility × Prec. Context	N/A	N/A	N/A
	Visibility × Foll. Context	N/A	N/A	N/A
	Visibility × First Mention	N/A	N/A	N/A

probably cannot be attributed to an effect of Visibility. The effect of Preceding Context appeared to disappear in Visible IDS, but it did not change according to Visibility here; and the effect of Unigram Probability weakened in Visible ADS compared to Non-Visible ADS, but did not change in the IDS/ADS regressions.

As with Analysis I, it is difficult to explain these results concisely with a spandrel account. We could again propose that an effect, this time of Unigram Probability, weakened in the Visible condition, because Visible dialogues are less cognitively demanding. However, in Analysis I, we saw that the effects of Local Context disappeared for the putatively "easier" condition, while there was no change in the effect of Unigram Probability. For this account to go through, then, we would need one version of it that explains why there is a floor effect in production difficulty according to Local Context when the conversation topic is easy, and a different version that explains why there is a floor effect in production difficulty according to Unigram Probability when the listener is visible. Again, the spandrel view demands multiple accounts.

However, these results cohere under the efficiency account, because they suggest that these talkers provided less redundancy overall in the Visible condition compared to the Non-Visible condition. As the Visible condition is the condition with higher channel capacity, this result suggests that talkers provide more acoustic redundancy to compensate for a lower channel capacity. Secondarily, we have established that the differences between ADS and IDS cannot be attributed solely to visibility. Together, these results support a close relationship between predictability effects and communicative efficiency.

5. General Discussion

In our first analysis, we compared the predictability effects in speech directed to competent adult native speakers to speech directed to prelinguistic infants, a difference of listener characteristics, and found reliable differences according to listener type. In our second analysis, we compared the predictability effects that arise when speech is the only communication channel to the predictability effects that arise when there is an additional, specifically visual, communication channel. In this analysis, we found that the amount of phonetic redundancy falls measurably when channel capacity increases. These results suggest that talkers modulate predictability effects according to at least very coarse, salient, and easy-to-track listener and channel characteristics, and that talkers can consider these characteristics in not just a high-level utterance planning phase but during low-level articulation. This modulation is a natural prediction of efficiency accounts (Aylett and Turk, 2004; Jaeger, 2010), but requires multiple independent spandrels under a non-adaptive account. Thus, predictability effects probably reflect a tendency towards efficiency in an information-theoretic sense, even if some talker-centric predictability effects might result from how neural hardware works.

5.1. Different notions of efficiency

The information-theoretic view of efficiency that we advocate here is not the only efficiency-based account of predictability effects. In particular, Ferreira and Dell (2000) also appealed to efficiency in their explanation of these effects. However, their notion of efficiency is different from the information-theoretic view, and contrasts two different kinds of efficiency—efficiency "for the talker" and efficiency "for the listener." To illustrate, Ferreira and Dell focused on the case of optional function words (see also Ferreira, 2008; Roland et al., 2006). For example, in English, the sentential complementizer *that* can often be omitted:

1. The coach knew you missed practice.
2. The coach knew that you missed practice.

Ferreira and Dell proposed two conflicting notions of efficiency that might drive talkers' decisions. First, "Availability Based Production" proposes that talkers choose between alternative forms by pronouncing the one that is available first. If a talker can access *you* immediately after saying *The coach knew*, then the talker pronounces it immediately (in "the principle of immediate mention," Ferreira and Dell, 2000, p. 299), and omits *that*. If, on the other hand, a talker cannot access *you*, they say *that* as "a grammatical 'um'" to give their access system enough time to retrieve *that*. Ferreira and Dell interpret this as a strategy that is efficient "for the talker," because it minimizes production effort: pronouncing a word as soon as it is available is presumably easier than accessing it and keeping it in storage for later pronunciation.

Second, "Ambiguity-Sensitive Production" points out that sentences are sometimes locally ambiguous without the complementizer: the *you* of "The coach knew you" could be either a direct object or the subject of a sentential complement, while the *I* of "The coach knew I" could only be the subject of a sentential complement. Under Ambiguity-Sensitive Production, talkers pronounce complementizers in only those sentences that would be locally ambiguous without one. Ferreira and Dell interpret this strategy as efficient "for the listener," because it reduces the amount of ambiguity a listener must deal with.

With these two notions of efficiency in hand, Ferreira and Dell propose that language production reflects a trade-off between efficiency for the listener and efficiency for the talker, and conclude from a series of experiments (also including Ferreira, 2003) that talkers do not consider potential ambiguities, instead pronouncing "that" only to give access mechanisms enough time to provide the next word.

For the purposes of noisy-channel coding, however, we can see that what Availability Based Production calls "efficiency for the talker" is not communicative efficiency. Optional *that* is a form of redundancy, since it makes the signal longer without changing the message, and so, from an information-theoretic point of view, serves the purpose of guarding against a communicative error in channel coding (Levy and Jaeger, 2007; Jaeger, 2010). If omitting *that* allows a talker to pronounce a subject immediately but leads to a communication error, and the talker must speak again only to clarify the intended interpretation, then omitting *that* was troublesome for both the talker *and* the listener. Moreover, clarifications will probably involve signals that are longer than the original *that* because of the overhead of complying with grammaticality constraints: saying *that* in at-risk sentences provides redundancy much more cleanly than a clarification with the overhead, due to grammaticality, of starting with "No, I meant..." or similar. The "principle of immediate mention" is thus at most about very short-term ease for the talker; it is not an efficiency principle.

While Ferreira and Dell (2000) and Ferreira (2003) do not find evidence that talkers pronounce redundant *that* to avoid ambiguity, this result only holds when sentences are coded as "ambiguous" or "not ambiguous," a categorical notion of ambiguity that does not fit well with an information-theoretic approach. Studies that rely on a gradient, statistical notion of ambiguity, such as the bias of a verb toward Direct Object or Sentential Complement arguments, do find an effect of ambiguity on *that*-mentioning and other forms of syntactic redundancy (Jaeger, 2006, 2010, 2011; Wasow et al., 2011). Additionally, while Jaeger (2006) and Jaeger

(2010) do find effects of availability on that-mentioning, the magnitude of the availability effects is much smaller than the effect of predictability (Jaeger, 2010, p. 35). Ferreira (2008) attributed this to a practice effect (e.g. Sentential Complement structures are harder to access for a verb that rarely takes Sentential Complements), but Jaeger (2013) pointed out that other similar phenomena (such as predictability-sensitive contraction, Frank and Jaeger, 2008) cannot be understood as a practice effect.

5.2. Optimal or closer to optimal?

It should be noted that predictability effects reflect a tendency *towards* efficient codes. It is not at this point clear that talkers actually achieve the optimal communication rate. In fact, there are reasons to think it is unlikely that they do. Ferrer-i-Cancho et al. (2013) showed that, when communicating at the channel capacity, signal characters are uniformly distributed and independent. We can view language as having different notions of signal 'character' (e.g., phones, words, or something else), but all of the obvious options are neither uniformly distributed nor independent. Moreover, this result implies that optimal codes cannot be decoded incrementally, but humans do decode language incrementally (Altmann and Kamide, 1999; Ito and Speer, 2008; Sedivy et al., 1999; Tanenhaus et al., 1995, 1996). Incremental decoding relies on reliably associating sequences of signal characters with the same string of meaning characters across signals. However, components of incremental linguistic meanings are not themselves uniformly and independently distributed, so a reliable association between the same sequence of message characters and the same sequence of signal characters would produce correlations between signal characters.

Another way of understanding this point about incrementality begins by noting that optimal encoding is possible only by encoding an entire *sequence* of message characters at once (using a *block code*) and in fact the proofs of optimality rely on encoding increasingly long sequences of message characters at once.[8] If we encode sequences of message characters, then the signal for one sequence may be completely different from the signal for a very similar sequence. Such sequences must be encoded all at once, and also decoded all at once, after observing the entire signal, rather than incrementally. To see why block coding is necessary to approach optimality, remember from the Background section that the optimal code length for a message is its negative log probability. This quantity will almost never be an integer, and so can be approached only by

[8] Arithmetic codes can be decoded incrementally while approaching optimality. However, arithmetic codes use a completely different signal for the same message character in different parts of a message. Natural languages do not work like this. For example, a word's phonological form at the beginning of a sentence is largely the same as its form later in the sentence. If a code consistently uses similar signals for the same component of a message, then block coding is necessary to approach theoretical limits.

encoding multiple message characters at the same time. The same reasoning holds for channel coding, if we view the "message characters" of channel coding to be the signal characters of the source code. Optimal codes thus differ from human language in (at least) two important ways; the signal characters are uniformly and independently distributed, and signals cannot be decoded incrementally.

One interesting point of speculation is that linguistic structure might result from a trade-off over evolutionary time between block-coding and incremental processing. Under this scenario, as our pre-linguistic ancestors sought to communicate longer messages, the one-to-one symbol-meaning mapping of most non-linguistic communication would be increasingly inadequate, introducing pressure towards information-theoretically efficient communication. However, existing pressures for immediate interpretation would preclude fully optimal block codes that cannot be decoded until the end of the utterance. These conflicting pressures were adjudicated by developing a hierarchy of block codes operating over different time scales: a block code for an entire turn would be optimal, but an amalgam of inter-related block codes for morphological paradigms, syntactic categories, and so on, is pretty good. If this speculation is on the right track, communicative pressures, in two fundamental and conflicting forms, would be central to the evolution of linguistic structure rather than, as advocated by Berwick et al. (2013), "ancillary".

In any case, the results presented above providing an important new view that the information-theoretic framing of efficiency accounts provides the best explanation for predictability effects.

5.3. Future work

Our first analysis found that talkers considered listener characteristics in producing predictability effects. While previous work had indicated that fine-grained listener characteristics may be too expensive to consider during low-level articulation, our analysis indicated that coarser listener characteristics can be cheap enough to consider during low-level articulation. One obvious future direction is to look at the ability of talkers to consider less dramatic listener characteristics. For example, Uther et al. (2007) found that talkers produced overall more redundant speech, with hyperarticulated vowels, when addressing foreigners. This suggests some degree of adaptation of channel coding to foreignness, with talkers assuming a higher base-rate of noise, and it would be interesting to see if talkers also adjust source-coding to foreignness.

Our second analysis found that talkers took advantage of the increased visual channel, but it remains unclear how talkers used that channel. As mentioned in the introduction, Boyle et al. (1994) found that pairs sought to establish eye contact during periods of difficulty; perhaps the visual channel simply allowed talkers to identify when they had neglected to provide enough redundancy in advance.

It is also unclear how the results from analysis 2 generalize to more realistic situations. The visual arrangement of the Map Task set-up is fairly unnatural with the most useful visual information (the partner's map) hidden from view. More realistic studies on the use of the visual channel to communicate information would investigate the role of gesture, for example, or the role of facial expressions.

5.4. Conclusion

Information theory provides a well-understood mathematical formalization for understanding language behavior as a communicative endeavor. In this article, we have explored the implications of information theory more fully, and shown how different kinds of predictability effects serve communicative goals in different ways. Moreover, while some kinds of predictability effects are consistent with non-optimizing access processes, others are more difficult to explain under this account, yet are closely tied to efficient communication. We examined two kinds of communication-oriented predictability effects, and found evidence that talkers adjust predictability effects according to listener characteristics and to the communicative scenario. Our more detailed information-theoretic framing along with these results bolster the view that a drive towards communication efficiency shapes language behavior, and fuels speculation that communicative pressures may have precipitated linguistic structure itself.

References

Altmann, G.T., Kamide, Y., 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition 73, 247–264.

Aylett, M., Turk, A., 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech 47, 31–56.

Baayen, R., Davidson, D., Bates, D., 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59, 390–412.

Baese-Berk, M., Goldrick, M., 2009. Mechanisms of interaction in speech production. Language and Cognitive Processes 24, 527–554.

Bard, E.G., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., Newlands, A., 2000. Controlling the intelligibility of referring expressions in dialogue. Journal of Memory and Language 42, 1–22.

Bard, E.G., Aylett, M.P., 2005. Referential Form, Word Duration, and Modeling the Listener in Spoken Dialogue. MIT Press, Cambridge, MASS. pp. 173–191.

Beckman, M., Hirschberg, J., Shattuck-Hufnagel, S., 2005. The original ToBI system and the evolution of the ToBI framework, in: Jun, S.A. (Ed.), Prosodic Typology–The Phonology of Intonation and Phrasing. Oxford University Press.

Bell, A., Brenier, J.M., Gregory, M., Girand, C., Jurafsky, D., 2009. Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language 60, 92–111.

Berwick, R.C., Friederici, A.D., Chomsky, N., Bolhuis, J., 2013. Evolution, brain, and the nature of language. Trends in Cognitive Sciences 17, 89–98.

Boyle, E.A., Anderson, A.H., Newlands, A., 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. Language and Speech 70.

Brent, M.R., Siskind, J.M., 2001. The role of exposure to isolated words in early vocabulary development. Cognition 81, 31–44.

Bresnan, J., Cueni, A., Nikitina, T., Baayen, R.H., 2007. Predicting the dative alternation. Royal Netherlands Academy of Science, Amsterdam. pp. 69–94.

Brown, P., Dell, G.S., 1987. Adapting production to comprehension: The explicit mention of instruments. Cognitive Psychology 19, 441–472.

Buz, E., Jaeger, T.F., Tanenhaus, M.K., 2014. Contextual confusability leads to targeted hyperarticulation, in: Proceedings of 36th annual meeting of the Cognitive Science Society.

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D., 2010. The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language Resources and Evaluation 44, 387–419. doi:10.1007/s10579-010-9120-1.

Clark, H.H., 1996. Using Language. Cambridge University Press, Cambridge, UK.

Clark, H.H., Carlson, T.B., 1982. Hearers and speech acts. Language 58, 332–373.

Clark, H.H., Marshall, C.R., 1981. Definite reference and mutual knowledge. Cambridge University Press, Cambridge, UK. pp. 10–63.

Coco, M.I., Keller, F., 2010. Scan pattern in visual scenes predict sentence production, in: Proceedings of the 32nd Annual Conference of the Cognitive Science Society.

Cook, S.W., Jaeger, T.F., Tanenhaus, M.K., 2009. Producing less preferred structures: More gestures, less fluency, in: Proceedings of the 31st Annual Conference of the Cognitive Science Society.

Dell, G.S., Brown, P.M., 1991. Mechanisms for Listener-Adaptation in Language Production: Limiting the Role of the "Model of the Listener". Psychology Press. pp. 105–129.

Esteve-Gibert, N., Prieto, P., 2013. Prosodic structure shapes the temporal realization of intonation and manual gesture movements. Journal of Speech, Language, and Hearing Research 56, 850–864.

Fernald, A., Simon, T., 1984. Expanded intonation contours in mothers' speech to newborns. Developmental Psychology 40, 104–113.

Ferreira, V.S., 2003. The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. Journal of Memory and Language 48, 209–246.

Ferreira, V.S., 2008. Ambiguity, accessibility, and a division of labor for communicative success. Psychology of Learning and Motivation: Advances in Research and Theory 49, 209–246.

Ferreira, V.S., Dell, G.S., 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. Cognitive Psychology 40, 296–340.

Ferrer-i-Cancho, R., 2005. Zipf's law from a communicative phase transition. The European Physical Journal B - Condensed Matter and Complex Systems 47, 449–457. URL: http://dx.doi.org/10.1140/epjb/e2005-00340-y, doi:10.1140/epjb/e2005-00340-y.

Ferrer-i-Cancho, R., Dębowski, Ł., Moscoso del Prado Martín, F., 2013. Constant conditional entropy and related hypotheses. Journal of Statistical Mechanics: Theory and Experiment 7. doi:10.1088/1742-5468/2013/07/L07001.

Fine, A.B., Frank, A.F., Jaeger, T.F., Van Durme, B., 2014. Biases in predicting the human language model, in: Proceedings of the Association for Computational Linguistics, pp. 7–12.

Frank, A., Jaeger, T.F., 2008. Speaking rationally: Uniform information density as an optimal strategy for language production, in: Proceedings of CogSci, pp. 933–938.

Fukumura, K., van Gompel, R.P.G., 2012. Producing pronouns and definite noun phrases: Do speakers use the addressee's discourse model? Cognitive Science 36, 1289–1311.

Gahl, S., Garnsey, S.M., 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. Language 80, 748–775.

Gahl, S., Garnsey, S.M., Fisher, C., Matzen, L., 2006. "That sounds unlikely": Syntactic probabilities affect pronunciation, in: Proceedings of the 27th meeting of the Cognitive Science Society.

Gahl, S., Yao, Y., Johnson, K., 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. Journal of Memory and Language 66, 789–806.

Galati, A., Brennan, S.E., 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? Journal of Memory and Language 62, 35–51.

Gould, S.J., Lewontin, R.C., 1979. The spandrels of san marco and the panglossian paradigm: A critique of the adaptationist programme. Proceedings of the Royal Society of London. Series B. Biological Sciences 205, 581–598. doi:10.1098/rspb.1979.0086.

Graf, H.P., Cosatto, E., Strom, V., Huang, F.J., 2002. Visual prosody: Facial movements accompanying speech, in: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition.

Grice, H.P., 1975. Logic and Conversation. Academic Press, New York. volume III: Speech Acts. pp. 41–58.

Harrell, F.E., 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York.

Heller, D., Gorman, K.S., Tanenhaus, M.K., 2012. To name or to describe: Shared knowledge affects referential form. Topics in Cognitive Science 4, 290–305.

Horton, W.S., Gerrig, R.J., 2002. Speakers' experiences and audience design: knowing *when* and knowing *how* to adjust utterances to addressees. Journal of Memory and Language 47, 589–606.

Horton, W.S., Gerrig, R.J., 2005. The impact of memory demands on audience design during language production. Cognition 96, 127–142.

Horton, W.S., Keysar, B., 1996. When do speakers take into account common ground? Cognition 59, 91–117.

Ito, K., Speer, S., 2008. Anticipatory effects of intonation: Eye movements during instructed visual search. Journal of Memory and Language 58, 541–573.

Jaeger, T.F., 2006. Redundancy and Syntactic Reduction in Spontaneous Speech. PhD dissertation. Stanford University.

Jaeger, T.F., 2010. Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology 61, 23–62.

Jaeger, T.F., 2011. Corpus-based Research on Language Production: Information Density and Reducible Subject Relatives. CSLI Publications, Stanford. pp. 161–197.

Jaeger, T.F., 2013. Production preferences cannot be understood without reference to communication. Frontiers in Psychology 4, 247–264.

Jaeger, T.F., Ferreira, V., 2013. Seeking predictions from a predictive framework. Behavioral and Brain Sciences 36, 359–360.

Claude Junqua, J., 1996. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. Speech Communciation 20, 13–22.

Khouw, E., Ciocca, V., 2007. Perceptual correlates of Cantonese tones. Journal of Phonetics 35, 104–117.

Kirov, C., Wilson, C., 2012. The specificity of online variation in speech production, in: Proceedings of the 34th annual meeting of the Cognitive Science Society, pp. 587–592.

Kuperman, V., Bresnan, J., 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. Journal of Memory and Language 66, 359–383.

Kurumada, C., Jaeger, T.F., 2013. Communicatively efficient language production and case-marker omission in japanese, in: Proceedings of 35th annual meeting of the Cognitive Science Society.

Kurumada, C., Jaeger, T.F., To Appear. Communicative efficiency in language production: Optional case-marking in japanese .

Levy, R., Jaeger, T.F., 2007. Speakers optimize information density through syntactic reduction, in: Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems.

Lindblom, B., 1990. Explaining phonetic variation: A sketch of the H & H theory. Kluwer Academic Publishers. pp. 403–439.

Lockridge, C.B., Brennan, S.E., 2002. Addressees' needs influence speakers early syntactic choices. Psychonomic Bulletin and Review 9, 550–557.

Lombard, E., 1910a. Note rectificative. Ann. Mal. Oreil. Larynx 36, 34–35.

Lombard, E., 1910b. A propos de la note rectificative et de la rectification. Ann. Mal. Oreil. Larynx 36, 111–112.

Lombard, E., 1911. Le signe de l'élévation de la voix. Ann. Mal. Oreil. Larynx 37, 101–119.

Manin, D., 2006. Experiments on predictability of word in context and information rate in natural language. Journal of Information Processes 6, 229–236.

Martin, A., Utsugi, A., Mazuka, R., 2014. The multidimensional nature of hyperspeech: Evidence from japanese vowel devoicing. Cognition 132, 216–228.

Mathews, D., Lieven, E., Theakston, A., Tomasello, M., 2006. The effect of perceptual availability and prior discourse on young children's use of referring expressions. Applied Psycholinguistics 27, 403–422.

Papoušek, M., Hwang, S.C., 1991. Tone and intonation in mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction. Applied Psycholinguistics 12, 481–504.

Peramunage, D., Blumstein, S.E., Myers, E.B., Goldrick, M., Baese-Berk, M., 2011. Phonological neighborhood effects in spoken word production: An fMRI study. Journal of Cognitive Neuroscience 23, 527–554.

Piantadosi, S., Tily, H., Gibson, E., 2011. Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences 108, 3526. URL: `http://colala.bcs.rochester.edu/papers/PNAS-2011-Piantadosi-1012551108.pdf`.

Pickett, J.M., 1956. Effects of vocal force on the intelligibility of speech sounds. Journal of the Acoustic Society of America 28, 902–905.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E., 2007. Buckeye corpus of conversational speech (2nd release) [www.buckeyecorpus.osu.edu].

Priva, U.C., 2008. Using information content to predict phone deletion, in: Proceedings of the 27th west coast conference on formal linguistics, pp. 90–98.

Roche, J., Dale, R., Kreuz, R.J., 2010. The resolution of ambiguity during conversation: More than mere mimicry?, in: Proceedings of 32nd annual meeting of the Cognitive Science Society.

Roche, J.M., Paxton, A., Ibarra, A., Tanenhaus, M.K., 2014. From minor mishap to major catastrophe: Lexical choice in miscommunication, in: Knauff, M., Pauen, M., Sebanz, N., Wachsmuth, I. (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society.

Roland, D., Elman, J.L., Ferreira, V.S., 2006. Why is *that*? Structural prediction and ambiguity resolution in a very large corpus of English sentences. Cognition 98, 209–246.

Rytting, C.A., Brew, C., Fosler-Lussier, E., 2010. Segmenting words from natural speech: subsegmental variation in segmental cues. Journal of Child Language 37, 513–543.

Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., Carlson, G.N., 1999. Achieving incremental semantic interpretation through contextual representation. Cognition 71, 109–147.

Seyfarth, S., 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. Cognition 133, 140–155.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423.

Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A., 1988. Effects of noise on speech production: Acoustic and perceptual analyses. Journal of the Acoustic Society of America 84, 917–928.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K., Sedivy, J.C., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632–1634.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C., 1996. Using eye-movements to study spoken language comprehension: Evidence for visually-mediated incremental interpretation. MIT Press, Cambridge, MA. pp. 457–478.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., Bresnan, J., 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. Language and Cognition 1, 147–165.

Uther, M., Knoll, M.A., Burnham, D., 2007. Do you speak E-N-G-L-I-S-H? A comparison of foreigner- and infant-directed speech. Speech Communication 49, 2–7.

Vance, T.J., 1977. Tonal distinction in Cantonese. Phonetica 34, 93–107.

Wasow, T., Jaeger, T.F., Orr, D.M., 2011. Lexical variation in relativizer frequency. De Gruyter. pp. 175–195.

Werker, J.F., Pegg, J.E., McLeod, P.J., 1994. A cross-language investigation of infant preference for infant-directed communication. Infant Behavior and Development 17, 323–333.

Whitehill, T.L., Ciocca, V., Chow, T.Y., 2000. Acoustic analysis of lexical tone contrast in dyarthria. Journal of Medical Speech and Language Pathology 8, 337–344.

Zhao, Y., Jurafsky, D., 2009. The effect of lexical frequency and Lombard reflex on tone hyperarticulation. Journal of Phonetics 37, 231–247.

Zipf, G.K., 1949. Human behavior and the principle of least effort. Addison-Wesley.