The communicative function of ambiguity in language

Steven T. Piantadosi

Harry Tily

Edward Gibson
Department of Brain and Cognitive Sciences, MIT

Abstract

We present a general information-theoretic argument that all efficient communication systems will be ambiguous, assuming that context is informative about meaning. We also argue that ambiguity additionally allows for greater ease of processing by allowing efficient linguistic units to be re-used. We test predictions of this theory in English, German, and Dutch. Our results and theoretical analysis suggest that ambiguity is a functional property of language that allows for greater communicative efficiency. This provides theoretical and empirical arguments against recent suggestions that core features of linguistic systems are not designed for communication.

Introduction

Ambiguity is a pervasive phenomenon in language which occurs at all levels of linguistic analysis. Out of context, words have multiple senses and syntactic categories, requiring listeners to determine which meaning and part of speech was intended. Morphemes may also be ambiguous out of context, as in the English $-s$, which can denote either a plural noun marking (*trees*), a possessive (*Dylan's*), or a present tense verb conjugation (*runs*). Phonological forms are often mapped to multiple distinct word meanings, as in the homophones *too*, *two*, and *to*. Syllables are almost always ambiguous in isolation, meaning that they can be interpreted as providing incomplete information about the word the speaker is intending to communicate. Syntactic and semantic ambiguity are frequent enough to present a substantial challenge to natural language processing. The fact that ambiguity occurs on so many linguistic levels suggests that a far-reaching principle is needed to explain its origins and persistence.

The existence of ambiguity provides a puzzle for functionalist theories which attempt to explain properties of linguistic systems in terms of communicative pressures (e.g. Hock-

We would like to thank Mike Frank, Florian Jaeger, and Roger Levy for helpful discussions about this work. We thank three anonymous reviewers for suggesting improvements to this paper. This work was supported by a National Science Foundation graduate research fellowship and a Social, Behavioral & Economic Sciences Doctoral Dissertation Research Improvement Grant (to S.T.P.), and National Science Foundation Grant 0844472 (to E.G.).

ett, 1960; Pinker & Bloom, 1990). One might imagine that in a perfect communication system, language would completely disambiguate meaning. Each linguistic form would map bijectively to a meaning, and comprehenders would not need to expend effort inferring what the speaker intended to say. This would reduce the computational difficulties in language understanding and comprehension because recovering meaning would be no more complex than, for instance, compiling a computer program. The communicative efficacy of language might be enhanced since there would be no danger of comprehenders incorrectly inferring the intended meaning. Confusion about "who's on first" could not occur.

Indeed, the existence of ambiguity in language has been argued to show that the key structures and properties of language *have not* evolved for purposes of communication or use:

> The natural approach has always been: Is it well designed for use, understood typically as use for communication? I think that's the wrong question. The use of language for communication might turn out to be a kind of epiphenomenon. ... If you want to make sure that we never misunderstand one another, for that purpose language is not well designed, because you have such properties as ambiguity. If we want to have the property that the things that we usually would like to say come out short and simple, well, it probably doesn't have that property. (Chomsky, 2002, p107)

Here, we argue that this perspective on ambiguity is exactly backwards. We argue, contrary to the Chomskyan view, that ambiguity is in fact a *desirable* property of communication systems, precisely because it allows for a communication system which is "short and simple." We argue for two beneficial properties of ambiguity: first, where context is informative about meaning, unambiguous language is partly redundant with the context and therefore inefficient; and second, ambiguity allows the re-use of words and sounds which are more easily produced or understood. Our approach follows directly from the hypothesis that language approximates an optimal code for human communication, following a tradition of research spearheaded by Zipf which has recently come back into favor to explain both the online behavior of language users (e.g. Genzel & Charniak, 2002; Aylett & Turk, 2004; Jaeger, 2006; Levy & Jaeger, 2007, i.a.) and the structure of languages themselves (e.g. Ferrer i Cancho & Solé, 2003; Ferrer i Cancho, 2006; Piantadosi, Tily, & Gibson, 2011). In fact, our specific hypothesis is closely related to a theory initially suggested by Zipf (1949).

In Zipf's view, ambiguity fits within the framework of his grand unifying *principle of least effort* and could be understood by considering the competing desires of the speaker and the listener. The speaker can minimize their effort if all meanings are expressed by one simple, maximally ambiguous word, say, *ba*. To express a meaning such as "The accordion box is too small," the speaker would simply say *ba*. To say "It will rain next Wednesday," the speaker would say *ba*. Such a system is very easy for speakers since they do not need to expend any effort thinking about or searching memory to retrieve the correct linguistic form to produce. Conversely, from the comprehender's perspective, effort is minimized if each meaning maps to a distinct linguistic form, assuming that handling many distinct word forms is not overly difficult for comprehenders. In that type of system, the listener does not need to expend effort inferring what the speaker intended, since the linguistic signal would leave only one possibility.

Zipf suggested that natural language would strike a balance between these two opposing forces of *unification* and *diversification*, arriving at a middle ground with some but not total, ambiguity. Zipf argued this balance of speakers' and comprehenders' interests will be observed in a balance between frequency of words and number of words: speakers want a single (therefore highly frequent) word, and comprehenders want many (therefore less frequent) words. He suggested the balancing of these two forces could be observed in the relationship between word frequency and rank frequency: the vocabulary was "balanced" because a word's frequency multiplied by its frequency rank was roughly a constant, a celebrated statistical law of language[1]. Ferrer i Cancho and Solé (2003) provide a formal backing to Zipf's intuitive explanation, showing that the power law distribution arises when information-theoretic difficulty for speakers and comprehenders is appropriately balanced. Zipf (1949) further extends his thinking to the distribution of word meanings by testing a quantitative relationship between word frequency and number of meanings. He derives a *law of meaning distribution* from his posited forces of unification and diversification, arguing that the number of meanings a word has should scale with the square root of its frequency. Zipf reports a very close empirical fit for this prediction. Functionalist linguistic theories have also posited trade-offs between total ambiguity and perfect and unambiguous logical communication (e.g. Givón, 2009), although to our knowledge these have not been evaluated empirically.

Zipf's statement of the way ambiguity might arise as a trade-off between speaker and hearer pressures has certain shortcomings. As pointed out by Wasow, Perfors, and Beaver (2005), it is unlikely that a speaker's effort is minimized by a totally ambiguous language, since confusion means that the speaker may need to expend effort clarifying what was intended. We show below that ambiguity is beneficial in exactly those situations where no additional clarification is required. Second, Zipf's argument required the hypothetical forces of unification and diversification. Our argument shows how the utility of ambiguity can be derived without positing that speakers want to produce one single concise word, or that comprehenders want a completely unambiguous system. We argue that Zipf's basic intuition about ambiguity—that it results from a rational process of communication—is fundamentally correct. Instead of unification and diversification, we argue that ambiguity can be understood by the trade-off between two communicative pressures which are *inherent to any* communicative system: *clarity* and *ease*. A *clear* communication system is one in which the intended meaning can be recovered from the signal with high probability. An *easy* communication system is one which signals are efficiently produced, communicated, and processed. There are many factors which likely determine ease for human language: for instance, words which are easy to process are likely short, frequent, and phonotactically well-formed. Clarity and ease are opposed because there are a limited number of "easy" signals which can be used. This means that in order to assign meanings unambiguously or clearly, one must use words which are more difficult.

One example that illustrates this trade-off is the NATO phonetic alphabet. The NATO phonetic alphabet is the system of naming letters which is used by the military and pilots—A is "Alpha", B is "Bravo", C is "Charlie", etc. This system was created to avoid the confusion that might occur when one attempts to communicate similar-sounding letter

[1]See also Manin (2008), who derives the Zipfian distribution of word meanings by positing that languages evolve to avoid excessive synonymy.

names across a noisy acoustic channel. The way this was done was by changing letters to full words, adding redundant information so that a listener can recognize the correct letter in the presence of noise. The downside is that instead of letters having relatively short names, they have mostly bisyllabic full-word names—which take more time and effort to produce and comprehend—trading ease for clarity. Trade-offs in the other direction are also common in language: pronouns, for instance, allow speakers to refer to locally salient discourse entities in a concise way. They are ambiguous because they could potentially refer to anyone, but allow for greater ease of communication by being short and frequent, and potentially less difficult for syntactic systems (Marslen-Wilson, Levy, & Tyler, 1982; Ariel, 1990; Gundel, Hedberg, & Zacharski, 1993; Warren & Gibson, 2002; Arnold, 2008; Tily & Piantadosi, 2009).

Our approach complements previous work arguing that ambiguity is rarely harmful to communication in practice thanks to the comprehender's ability to effectively disambiguate between possible meanings (Wasow & Arnold, 2003; Wasow et al., 2005; Jaeger, 2006; Roland, Elman, & Ferreira, 2006; Ferreira, 2008; Jaeger, 2010). Additionally, several authors have previously discussed the possibility that ambiguity is a useful feature of language. Several cognitive explanations of ambiguity were discussed by Wasow et al. (2005). One is the possibility that ambiguity reduces the memory demands of storing a lexicon, though they conclude that human memory is probably not a bottleneck in vocabulary size. They also hypothesize that there may be some processing constraint against longer morphemes which leads to shorter morphemes being recycled for multiple meanings. This is one case of the theory we present and test in the next section: that forms are re-used when they are easy to process. Wasow et al. (2005) also suggest ambiguity might be useful in language contact situations, where speakers of both languages should ideally be able to handle words meaning two different things in two different situations. They also point out that ambiguity does sometimes serve a communicative function when speakers wish to be ambiguous intentionally, giving the example of a dinner guest who says "Nothing is better than your cooking" to express a compliment and an insult simultaneously. Neither of these arguments are especially compelling because it is unclear how they could explain the fact that linguistic ambiguity is *so* common.

Ambiguity has also been considered previously in the context of mathematical models of communication, as a capacity that may allow language to stay in a preferable portion of the model's parameter space Ferrer i Cancho (2006), or as a necessary factor in the development of a combinatorial communication system (Ferrer i Cancho & Loreto, in prep). In particular, Ferrer i Cancho and Loreto (in prep) discuss that combining linguistic units provides no advantage when each linguistic unit unambiguously communicates a full meaning, indicating that the compositional, combinatorial nature of language is intrinsically linked with the property that individual linguistic units (e.g. words) do not completely disambiguate a speaker's intended meaning. However, their analysis does not go so far as to say that entire combinatorial linguistic utterances should still be ambiguous, as we find with semantic or syntactic ambiguities that can only be resolved using contextual information.

An information-theoretic approach to ambiguity was pursued by Juba, Kalai, Khanna, and Sudan (2011), who argue that ambiguity allows for more efficient compression when speakers and listeners have boundedly different prior distributions on meanings. This complements the information-theoretic analysis we present in the next section, although study-

ing boundedly different priors requires a considerably more complex analysis.

The goal of the present paper is to develop an explanation for ambiguity which makes fewer assumptions than previous work, and is more generally applicable. Indeed, the explanations we present demonstrate that ambiguity is a desirable feature of *any* communicative system when context is informative about meaning. We argue that the generality of our results explains the pervasiveness of ambiguity in language, and shows how ambiguity likely results from ubiquitous pressure for *efficient communication*.

Two benefits of ambiguity

In this section we argue that efficient communication systems will be ambiguous when context is informative about what is being communicated. We present two similar perspectives on this point. The first shows that the most efficient communication system will not convey information already provided by the context. Such communication systems necessarily appear to be ambiguous when examined out of context. Second, we argue that specifically for the human language processing mechanisms, ambiguity additionally allows re-use of "easy" linguistic elements—words that are short, frequent, and phonotactically high probability.

Both these perspectives assume that disambiguation is not prohibitively costly (see Levinson, 2000)—that using information from the context to infer which meaning was intended does not substantially impede comprehension. We return to this issue in the discussion. We note here that our explanations for ambiguity do not prove ambiguity *necessarily* makes language more efficient. One could always construct an ambiguous linguistic system which was not efficient—for instance, one which leaves out information other than what is provided in the context, or re-uses particularly difficult linguistic elements. Instead, these benefits of ambiguity suggest that any system which strives for communicative or cognitive efficiency will naturally be ambiguous: ambiguity is not a puzzle for communicative theories of language.

Ambiguity in general communication

In this section, we motivate an information-theoretic view of ambiguity. We will assume that there exists a set M of possible meanings. For generality, we will allow M to range over any possible set of meanings. For instance, M might be the space of compositional semantic structures, the space of parse trees, or the set of word senses. The argument in this section is general to any space of meaning.

Intuitively, a linguistic form is ambiguous if it can map to more than one possible meaning. For instance, the word "run" is ambiguous because it can map to a large number of possible meanings, including a run in a pantyhose, a run in baseball, a jog, to run, a stretch of consecutive events, etc. It turns out, however, that we do not need to consider the ambiguity of specific words or linguistic units to argue that ambiguity is in general useful. This is because language can fundamentally be viewed as conveying bits of information about the speaker's intended meaning. By formalizing a notion of uncertainty about meaning, one can show that the optimally efficient communication system will look ambiguous, as long as context is informative about meaning.

We quantify the uncertainty that listeners would have about intended meaning by using *Shannon entropy*[2]. Shannon entropy measures the amount of information required on average to disambiguate which meaning in M is intended and is given by

$$\mathbf{H}[M] = -\sum_{m \in M} P(m) \log P(m), \tag{1}$$

where $P(m)$ is the probability that meaning m is the intended meaning. Shannon entropy quantifies information on a scale of *bits*. When $P(m) = 1$ for some m, no information about the meaning needs to be transmitted (since the intended meaning can always be guessed correctly without any communication) so the entropy is 0. Conversely, when the entropy is high, more bits of information are needed to disambiguate which of the possible meanings was intended. If we consider only two possible meanings, there is maximal uncertainty when both meanings are equally likely. In this case, we need exactly one bit of information to disambiguate which meaning was intended. This can be checked by plugging in $P(m_1) = P(m_2) = \frac{1}{2}$ into equation 1 above, to get 1 bit of uncertainty[3]. When one meaning is much more frequent than other, it requires less than less than 1 bit of information on average to disambiguate.

The notion of ambiguity in Equation 1 does not take into account context—only the listener's a-priori uncertainty about intended meaning. However, actual language use takes place with reference to world and linguistic context. Knowing that the speaker is playing baseball, for instance, will change the expectations of what meaning of "run" is intended. This means that the probability distribution $P(m)$ may depend on context, and therefore the Shannon entropy does as well. For convenience we will wrap all extra-linguistic factors, including discourse context, world context, world knowledge, etc. into a variable C, for "the context." We can then include C into the information-theoretic framework by measuring the entropy of M, conditioned on C:

$$\mathbf{H}[M \mid C] = -\sum_{c \in C} P(c) \sum_{m \in M} P(m \mid c) \log P(m \mid c). \tag{2}$$

Here, the rightmost sum is simply the entropy over meanings in the particular context $c \in C$. This part of the equation is the same as Equation (1), except that $P(m)$ has been replaced by the probability of m in context c, denoted $P(m \mid c)$. This entropy is weighted by a distribution $P(c)$ on contexts, meaning that $\mathbf{H}[M \mid C]$ can be interpreted as the expected entropy over meanings, in context.

While these equations provide ways to theoretically compute the entropy or ambiguity left by a linguistic element, what is more important is the relationship between these two entropy measures. In particular, if C is informative about meaning, then it is provably true (see Cover & Thomas, 2006, p20-30) that

$$\mathbf{H}[M] > \mathbf{H}[M \mid C]. \tag{3}$$

In other words, when the context C is known and informative, it necessarily decreases the entropy. The strictness of this inequality comes from the fact that context provides some

[2]See Cover and Thomas (2006) for a mathematical overview of information theory, and MacKay (2003) for a technical introduction.

[3]For log base 2

information about meaning. As an example, listeners may have uncertainty about whether a word like "run" is intended to be a noun or a verb. Out of context, it may be somewhat difficult to guess, and so the word "run" is highly ambiguous out of context. In context, however, it is often clear: syntactic information, as well as discourse context, provide a considerable amount of meaning about which "run" is intended. For instance, "a run" typically disambiguates "run" to a noun, but "we run" typically disambiguates to a verb. Equation 3 states that on average it must be the case that the meanings can be conveyed with fewer bits of information when context is taken into account.

An optimally efficient communication system will attempt to be as efficient as possible, meaning that it will not convey unnecessary information. Since language use occurs in contexts C, the least amount of information that language can convey without being ambiguous in context is $\mathbf{H}[M \mid C]$. Because Equation 3 is a strict inequality, the amount of information efficient language should convey will always be less than the entropy out of context, $\mathbf{H}[M]$. Shannon's source coding theorem (Shannon, 1948) showed, in part, that it would be impossible to disambiguate $\mathbf{H}[M]$ bits of uncertainty with fewer than $\mathbf{H}[M]$ bits of information without making errors. This means that no matter what $\mathbf{H}[M \mid C]$ bits of information the linguistic system communicates, it will never be able to remove $\mathbf{H}[M]$ bits of uncertainty: an efficient communication system will never be able to disambiguate language out of context. This means that when the individual units of an efficient communication system are viewed outside of their typical contexts, they will look ambiguous.

Note that this argument is very general in that we made no assumptions about the linguistic system, or the distribution of contexts or meanings. Additionally, we did not even make assumptions about what the contexts or meanings actually were: the argument is general enough to apply to all situations in which context is informative about meaning, whatever context and meaning happen to be, and therefore applies to all levels of linguistic analysis. A key assumption that is required is that speakers and listeners have the same—or very similar—coding schemes (corresponding to similar probabilistic models of language and the world), and also the same ability to use contextual information to constrain the possible meanings.

Ambiguity and minimum effort

Here, we present a second argument that ambiguity is a desirable property of a linguistic system because it potentially allows for greater overall ease of processing. The results in this section extends the information-theoretic proof above to the case where code words (e.g. phonological forms) vary in their difficulty for language processing systems. The argument is similar in spirit to Zipf (1949)'s proposal described in the introduction which held that ambiguity can result from trading off ease of production and ease of comprehension. Our proposal differs in how "ease" is quantified, and is general enough to potentially include Zipf's ideas, as well as other properties of language which have been found to affect processing difficulty. This point does not necessarily entail that languages do use ambiguity in this efficient way; in the next section, we establish with corpus studies that they do.

Suppose that L is a set of linguistic units—for instance, words. Each element of L varies in its difficulty for the comprehension and production mechanisms: short words are easier, phonotactically well-formed words are easier, frequent words are easier, etc. Each possible meaning in M is mapped to a linguistic unit in L. Ambiguity allows multiple

meanings $m_1, m_2 \in M$ to be mapped to the same linguistic unit $l \in L$; conversely, an unambiguous linguistic system would map only one meaning to each element of L. As above, the proof that ambiguity is desirable in this setup will assume that context is informative about meaning, and disambiguation is not prohibitively costly. Here, it will be most useful to consider that some meanings m_1 and m_2 which are well-enough disambiguated by context that they can both be assigned to the same lexical item without significant chance of confusion. This is likely the case with many word-sense ambiguities, for instance a run in pantyhose and a run in baseball.

In an unambiguous linguistic system, m_1 is mapped to $l_1 \in L$ and m_2 is mapped to $l_2 \in L$, with $l_1 \neq l_2$. Suppose that l_1 is easier than l_2 according to any metric of effort—length, phonotactic well-formedness, neighborhood size, frequency, broader memory considerations, or some overall combination of these. If l_1 is easier than l_2, and m_1 and m_2 are well-disambiguated by context, then we can always create a linguistic system which is easier overall by mapping m_1 and m_2 both to l_1. This costs the same in terms of effort every time we communicate m_1, but saves effort every time we communicate m_2, relative to the unambiguous system, since l_1 is less effort than l_2. Of course, in real language there are likely many pairs or tuples of meanings which can be assigned to the same linguist forms. This argument is meant to simplify and show that as long as there are at least two meanings which are unlikely to occur in the same contexts, the linguistic system can be improved by introducing ambiguity.

This shows that under very weak conditions, an unambiguous linguistic system can always be made easier to use by preferentially re-using the "easy" linguistic units. Unlike the information-theoretic argument, this argument assumes specific details about the linguistic elements which are being communicated—some involve less effort than others. Additionally, this argument requires the assumption that disambiguation is not prohibitively costly—otherwise a language with ambiguity would not be "easier." We find this plausible because many different meanings occur in largely non-overlapping contexts (for instance, in syntactic category ambiguity). In addition, the inference involved in disambiguation does not appear to be especially costly. In many if not all communicative situations, speakers easily infer a rich set of pragmatic and social consequences of language use. Levinson (e.g. Levinson, 2000) has argued extensively that in all aspects of communication, much of the speaker's intent is not explicitly coded in language but inferred through the hearer's knowledge of likely intentions, conventions of interaction, and common sense knowledge about the world. Like Zipf, Levinson assumes that this situation results from a trade-off between hearer and speaker pressures, and moreover argues that human cognitive abilities will favor communication systems which are heavy on hearer inference and light on speaker effort.

> The essential asymmetry is: inference is cheap, articulation expensive, and thus the design requirements are for a system that maximizes inference. (Hence ... linguistic coding is is to be thought of less like definitive content and more like interpretive clue.) (Levinson, 2000, p29)

Our theory relies on exactly this point: hearers are good at disambiguating in context, and therefore any effort the speaker makes to express a distinction that could have been inferred is wasted effort.

Empirical evaluation of ambiguity and effort

In the previous section, we presented two closely related arguments that ambiguity allows for more efficient communication systems. Both assumed that information is typically present to resolve ambiguities, and that using this information is relatively "cheap." The first argument looked at ambiguity from the perspective of coding theory, arguing that when context is informative, any good communication system will leave out information already in the context. The second assumed that codewords differ in their difficulty, and argued that as long as there are some ambiguities that context can resolve, efficient communication systems will use ambiguity to make communication easier. We do not view these arguments as distinct alternatives, but rather as two complementary ways of understanding how ambiguity is useful. It is unclear how one might test the first, information-theoretic argument, since it is a mathematical demonstration that ambiguity should exist; it does not make predictions about language other than the presence of ambiguity. The second account, however, directly predicts that linguistic units which require "less effort" should be more ambiguous. This would be a hallmark of an efficient communication system, one which has harnessed ambiguity as a core functional component of language.

One simple and intuitive measure of effort is word frequency: words which are used more often are generally processed more quickly than infrequent words. Indeed, previous results such as Reder, Anderson, and Bjork (1974) and (Zipf, 1949, p30) have found relationships between frequency and the amount of ambiguity in a word. However, frequency is correlated with other measures—most notably length—and is confounded since words with more meanings should be useful in more contexts, and thus be used more frequently. To our knowledge, no work has systematically evaluated multiple measures of difficulty and looked for effects of each, controlling for the others.

In this section we empirically evaluate the prediction that ambiguity allows for re-use of efficient linguistic units by looking at homophony, polysemy, and the ambiguity about meaning of different syllables, in English, German, and Dutch. Our basic approach is to measure properties of words and syllables which should influence their ease of processing, and see if easier linguistic units are preferentially re-used in language. We investigated the influence of three simple and easy-to-measure properties of words: length, frequency, and phonotactic probability. Both frequency and length are known to influence, for instance, on-line language processing (e.g. Rayner, 1998; Demberg, 2010) with longer and lower-frequency words taking longer to process. The phonotactic predictability measure uses Markov model to quantify how phonetically probable a word is, given all other words in the language. Intuitively, words that are re-used through ambiguity should be very high probability in order to increase cognitive and articulatory ease. While we only examined these three predictors, our theory predicts that any other measure which increases processing ease should also increase ambiguity.

We use several different techniques to analyze the influence of these factors on ambiguity. Ideally, one would measure ambiguity using the entropy over meanings for a given linguistic form. Unfortunately, entropy is difficult to estimate without statistical bias (see Paninski, 2003); in fact, the amount by which an entropy estimate is biased depends on the sample size—in this case, token frequency—meaning that one might expect correlations with frequency simply because of estimation error. This means that using entropy over

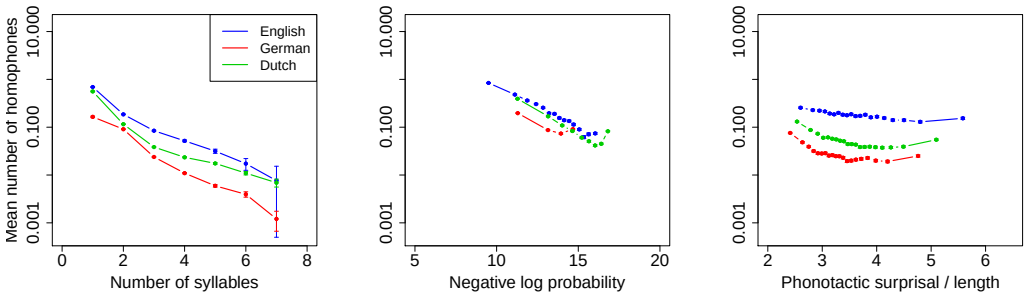
meanings as an outcome measure leads to results which are difficult to interpret. For this reason, we primarily present count data: for each linguistic unit, we count the number of possible meanings it has in order to measure its degree of ambiguity. There exist sophisticated regression techniques for counts, and the counts we use are hand-created, which means they should not be inherently correlated with other measures like frequency. We also present linear regression results using rank-ordered counts, a measure which makes less assumptions about the distribution of counts. Finally, for comparison we also include results using entropy measures over meanings, but caution that these are more difficult to interpret. All these analysis methods give nearly identical qualitative results, meaning that the findings we present are robust to analysis method and how ambiguity is quantified. We note that the three analysis methods are not statistically independent since the counts, rank-ordered counts, and entropy are all correlated.

Homophony

Here, we examine homophones in CELEX (Baayen, Piepenbrock, & Gulikers, 1995) and test the predictions of the previous section that phonological forms which are easier should be more ambiguous. For instance, the phonological form for "we" has a single homophone, "wee," since "we" and "wee" have different meanings but are pronounced the same. Here, words with multiple parts of speech—"experiment" the verb and "experiment" the noun—are also counted as homophones. Our prediction is that phonological forms which are high-frequency (low negative log probability), phonotactically well-formed, and short will map to many different word lemmas. We excluded words in CELEX containing spaces, hyphens, or apostrophes, leaving 65,417 English phonological forms, 310,668 German phonological forms, and 277,522 Dutch phonological forms. These mapped to a total of 77,243 English lemmas, 319,579 German lemmas, and 292,563 Dutch lemmas respectively.

Word length was measured by syllables, although measuring it by phonemes gives similar results. Word frequencies were taken from CELEX, and were transformed to negative log probabilities (unigram probabilities). We used add-one smoothing to prevent words from having zero frequencies, though similar results are found by excluding the lowest frequency words. To compute phonotactic surprisal, we used a simple triphone language model to compute the surprisal (negative log probability) of each phonological form. We trained the phonotactic language model using the word frequencies: for each phonological form w, we removed all but one token count of w from the corpus, and trained a language model on the *remaining* lexicon. This means the phonotactic model can be viewed as the probability of a phonological form training on all *other* words, smoothed with one token count of the current word. However, the results here are robust to the precise form of the phonotactic language model as biphone and quadphone models gave nearly identical results. The measure of phonotactic surprisal was divided by word length to prevent it from being collinear with length, and can therefore be interpreted as surprisal per phoneme, averaged over the entire word. With these covariates, multicollinearity was assessed by computing a *variance inflation factor*, which was below 2.0 in all languages, indicating a low-degree of collinearity, despite correlations between length and frequency (e.g. Zipf, 1936). Indeed, residualization of length on frequency gives identical qualitative results.

Figure 1 shows the mean number of homophones a given phonological form has. Each phonological form maps to at least one word in CELEX, and this figure shows the num-



(a) (b) (c)

Figure 1. : Number of additional meanings each a phonological form has, as a function of length, negative log probability, and phonotactic surprisal. This shows predicted effects, with shorter, lower negative log probability (higher frequency), and lower phonotactic surprisal forms having more meanings. Note all y-axes are logarithmically spaced. Error bars show standard errors within each bin.

ber of *additional* words mapped to each phonological form, for number of syllables, binned log (base e) negative log probability, and binned phonotactic log probability, in the three languages. This figure shows several clear patterns. First, Figure 1a shows that shorter phonological forms are assigned more meanings. These patterns all hold across the three languages examined here, indicating that ease seems to be a robust and cross-linguistic predictor of how many meanings will be assigned to each phonological form. Figure 1b also shows that lower negative log probability (higher frequency) phonological forms tend to be mapped to many more word meanings than higher negative log probability (lower frequency) phonological forms, across all languages. This is somewhat difficult to interpret because phonological forms with more meanings should be seen more simply because they can be used in more situations. However, that interpretation predicts a linear relationship between number of meanings and frequency: all else being equal, a word with k meanings should be used k times more than a word with 1 meaning. This figure demonstrates a linear relationship between number of meanings and *log* frequency, corresponding to a super-linear relationship between number of homophones and frequency. We therefore argue such a relationship likely results from the ease of processing more frequent wordforms, rather than merely the fact that phonological forms with more meanings can be used in more situations. Figure 1c shows that as average phonotactic surprisal per phoneme increases, words also tend to have fewer meanings. This effect tends to level out, showing no differences between the highest surprisal words, or, in the case of Dutch slight increases with the highest phonotactic surprisal words. These effects may result from poorer estimation in the highest phonotactic surprisal words, which have the lowest frequency phonotactic trigrams. In general, though, these plots show the predicted trends for the majority of data, indicating that phonotactically easier—higher probability according to the rest of the lexicon—phonological forms are assigned more meanings.

We performed several different types of regression analysis on this data. This allowed us to test the statistical significance of the trends in Figure 1 and evaluate the performance

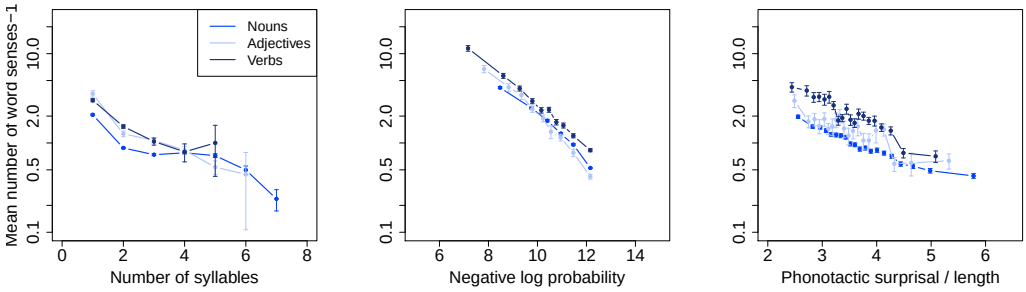
of each predictor while simultaneously controlling for effects of the other predictors. This is especially important in the case of, for instance, frequency and length, since these two variables are correlated and it is important to know that apparent effects of one variable do not result from correlations with another. For all regression analyses, we used standardized the covariates. We first used a quasi-Poisson regression to predict the number of additional meanings each phonological form in CELEX can map to (Gelman & Hill, 2006). This regression revealed significant effects of length, with longer phonological forms mapping to fewer words in all languages ($\beta > -0.85, t < -51.89, p < 0.001$ in each language). Higher negative log probability (lower frequency) words mapped to fewer meanings ($\beta > -0.71, t > 53.3, p < 0.001$ in each language). Second, words with higher phonotactic surprisal mapped to fewer words in German and Dutch ($\beta > -0.11, t < -6.59, p < 0.001$), but the trend was only marginally significant in the wrong direction in English ($\beta = 0.03, t = 2.01, p < 0.045$). This effect is non-significant if controlling for multiple comparisons. These results demonstrate that the tends in Figure 1 are statistically significant while controlling for other variables, with the exception of the English phonotactic curve. Several interactions were present, although they were generally of small magnitude and not of theoretical interest here.

We additionally performed a regression predicting a phonological form's *rank* in terms of number of meanings. Thus, the word with the most meanings was rank 1, the second most meanings was rank 2, etc. This revealed nearly identical qualitative results: predicted effects of longer words increasing rank ($\beta > 2436, t > 55.77, p < 0.001$ in each language), higher negative log probability (lower frequency) increasing rank ($\beta > 3175, t < -69.17, p < 0.001$ in each language), and higher phonotactic surprisal increasing rank ($\beta > 142, t > 3.2, p < 0.002$ in each language). Finally, for comparison, we also included a regression predicting the entropy over lemmas for each phonological form, as measured using maximum likelihood entropy estimation with the CELEX frequency counts. As discussed above, entropy is difficult to estimate, but the results here appear quite robust even with estimated entropies. This regression revealed significant, predicted effects: increasing length decreases entropy ($\beta < -0.0098, t < -52.2, p < 0.001$ in each language), higher negative log probability (lower frequency) decreases entropy ($\beta < -0.015, t > 63.3, p < 0.001$ in each language), and increasing phonotactic surprisal decreases entropy ($\beta < -0.0019, t < -9.9, p < 0.001$ in each language).

These regression analyses indicate that each factor that we predicted to increase ease of use, also increases the number of meanings assigned to a phonological form. This finding is robust to the way in which ambiguity is quantified.

Polysemy

Next, we consider similar predictions about the number of word *senses* each word has as a function of the word's length. For this analysis, we looked at word forms found in the English versions of WordNet (Fellbaum et al., 1998) and CELEX (Baayen et al., 1995). We chose to look at part of speech categories separately to ensure that findings are not driven by a single part of speech category, and also to check that these effects go beyond effects of homophony. For each word and part of speech, we computed the number of senses using WordNet. Words such as "run" have many *senses*—while homophone sets only distinguish substantially different meanings, word senses separate related meanings, such as those in



(a) (b) (c)

Figure 2. : Number of word senses as a function of a word's length, negative log probability, and phonotactic probability. This shows predicted effects, with shorter, higher frequency (lower negative log probability), and lower phonotactic surprisal forms having more senses. All y-axes are logarithmically spaced and error bars show standard errors within each bin.

"John runs to the store", *"she runs her fingers through her hair"*, and *"the train runs between Boston and New York"*. For each word, CELEX was used to find the phonological length of each word, as well as its phonotactic probability and frequency (negative log probability), using the same methods as the previous section. However, here we analyzed the number of senses assigned to each word lemma without collapsing by phonological form. This resulted in 20,582 nouns, 3175 verbs, and 1,536 adjectives, with 41,206, 10,358, and 3,770 total senses respectively. Multicollinearity was assessed by computing a variance inflation factor, which was below 1.84 in all languages, indicating a low-degree of collinearity; residualization of length on frequency yielded identical qualitative results.

Figure 2 shows predicted effects in each part of speech, and for each measure. This reveals the predicted trends in nearly the full range of all measures. Note that the lower negative log probability noun bins that do not match the qualitative patterns each contain only one word. For the majority of bins across the range of variables, factors which should increase ease also increase the number of word senses. The frequency results here can be compared to Zipf (1949, p27-30), who previously found that more frequent words have more meanings using a dictionary. He found a linear relationship with a slope close to -0.5, the theoretical slope according to his law of meaning distribution. Zipf presented these results as evidence for opposing forces of unification and diversification, which provided the basis for his explanation of ambiguity. The results here are similar, but with a different theoretical basis. We argue that frequent words—like phonotactically well-formed and short words—have more meanings because they are easier to process. In contrast, Zipf argued that frequent words have more meanings because such a relationship optimally balances concerns of speakers and listeners.

As with homophony above, our primary regression technique was a Poisson regression predicting the additional number of senses a word has, from each of the predictors. As above, this regression revealed longer words have fewer senses ($\beta < -.15, t < -4.1, p < 0.001$ for each part of speech category), higher negative log probability (lower frequency) words have fewer senses ($\beta < -0.47, t > 23.6, p < 0.001$ for each part of speech category), and higher

phonotactic surprisal have fewer senses ($\beta < -.15, t < -3.95, p < 0.001$ for each part of speech category). As above, a linear regression on rank number of senses revealed identical qualitative patterns. A regression on entropy was not performed because enough data on the frequency of various word senses was not available; however, given the results in the previous section, it is likely that these results would generalized to entropy-based ambiguity measures.

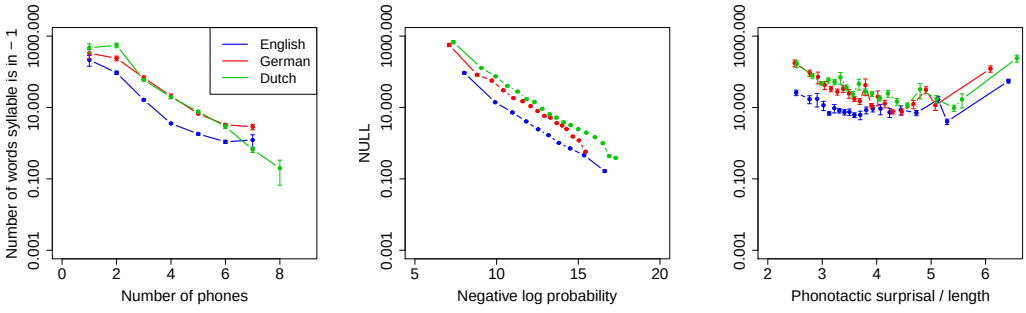
This finding replicates our results showing the relationship between processing difficulty and ambiguity using an alternative measure of ambiguity in English (the number of senses for each word). It also shows that the effect is not limited to homophones that differ in part-of-speech category.

Syllables

The previous two sections tested length-ambiguity relationships on two different analyses at the word level: homophony and polysemy. In this section we extend the analyses of the previous sections to the syllable level. Our prediction for syllables is identical: easier syllables should convey less information about meaning than harder syllables. In this case, we take syllables to be informative about the words that they appear in. Given just one syllable, one will have some incomplete information about what word the speaker is attempting to communicate. In this sense, individual syllables are ambiguous about meaning, and it is only when they are heard in the context of other syllables, words, and discourse situations that they can be used to unambiguously communicate meaning.

We analyzed syllables in words from CELEX. In each language we computed the number of words each syllable appears in. Syllable frequencies and phonotactic log probabilities were computed using the same procedure as the previous two sections. The length of each syllable was measured as the number of phones in its phonological transcription. We computed phonotactic log probability using a trigram model, and computed the negative log probability of each syllable according to its total token count in CELEX. This resulted in a total of 11,243, 10,816, and 11,979 syllable types in English, German and Dutch respectively. Here, we take syllables to be informative about the word they appear in: each syllable conveys information about the word the speaker is intending to say.

The relationship between the three "ease" measures and number of words a syllable appears in is shown in Figure 3. As in the previous sections, we subtracted one from each of these counts since each syllable must appear in at least one word to be in the sample. This shows that syllables pattern similarly to words, except in the case of phonotactic predictability. The syllables with lowest phonotactic surprisal do appear in the most words; however, very high phonotactic surprisal syllables also tend to appear in many words. This quadratic trend is significant using a quadratic term in a quasi-Poisson regression, for English and Dutch ($p < 0.001$), but not German ($p = 0.52$). We believe this trend is an artifact of our phonotactic surprisal model, which has increased estimation error for the high phonotactic surprisal (low phonotactic probability) phones. This interpretation is supported by the absence of a quadratic trend using a two-phone model ($p > 0.25$ for each language). Alternatively, it may be the case that other articulatory effects (e.g. the Obligatory Contour Principle: Frisch, Broe, & Pierrehumbert, n.d.; McCarthy, 1985, 1986; Graff & Jaeger, 2009) are present at the syllable level and that this trend results from other kinds of articulatory constraints.



(a) (b) (c)

Figure 3. : Number of words a syllable appears in, as a function of the syllable's length, negative log probability, and phonotactic surprisal. This shows predicted effects of length and negative log probability, with a more complex pattern in phonotactic surprisal. All y-axes are logarithmically spaced and error bars show standard errors within each bin.

A quasi-Poisson regression shows all effects in the predicted direction for English and Dutch: increasing length decreases the number of words a syllable appears in ($\beta < -0.07, t < -2.6, p < 0.01$ in each language), higher negative log probability (lower frequency) decreases the number of words ($\beta < -1.38, t > 65.3, p < 0.001$ in each language), and increasing phonotactic surprisal tends to decrease the number of words ($\beta < -0.18, t < -4.6, p < 0.001$ in each language). In German, effects of frequency ($\beta = -1.74, t = -60, p < 0.001$) and phonotactic surprisal ($\beta = -.16, t = -2.88, p < 0.01$) were significant in the predicted directions; length, however, was significant in a non-predicted direction ($\beta = 0.21, t = 4.83, p < 0.001$). Multicollinearity was assessed by computing a variance inflation factor, which was below 3.2 in all languages, indicating a low-degree of collinearity. However, residualization of length on frequency yielded the predicted effects, reverse effects, and null effects in English, German, and Dutch respectively. This indicates that the observed effects of length are only independent of frequency in English, and that length effects might not be present in all languages.

Regressions predicting rank-order number of words a syllable appears in yielded patterns identical to the quasi-Poisson regression; regressions predicting entropy yielded a null result for German syllable length (rather than a significant result the wrong way) and predicted directions for all other languages and variables.

In general, these results indicate that generally the predictors of ease extend to syllable units, although not in the case of German syllable length. As discussed above, it is likely that at the syllable level, other kinds of constraints such as articulation exert a stronger influence on the design of lexical systems. In general, this syllable analysis is interesting in part because syllables are not generally taken to be ambiguous in the same way that words or sentences are. Syllables are not, on their own, meaningful units of language, and therefore it may seem strange to describe them as ambiguous. However syllables are informative about intended meanings. In this case, we take the intended meaning to be the word lemma which being communicated, the same unit of meaning used in the homophony analysis. The results show that syllables are differentially informative about this intended

meaning, and that the same factors which influence processing ease on a word-level are also seen at the sub-word, or indeed sub-meaning, level of syllables.

General Discussion

We have presented two related arguments that show a well-designed communication system will be ambiguous, when examined out of context. We tested predictions of this theory, showing that words which are more efficient are preferentially re-used in language through ambiguity, allowing for greater ease overall. Our regression on homophones, polysemous words, and syllables—though similar—are theoretically and statistically independent. We therefore interpret positive results in each as strong evidence for the view that ambiguity exists for reasons of communicative efficiency. We note, however, that the languages tested are historically-related, meaning that further work will be needed to establish stronger typological generalizations.

Our analyses used regressions, which means that coefficients can be interpreted as the effect of one covariate, while controlling for others. This is important because if one finds a relationship between, say, ambiguity and length, it is important to show that this apparent effect is not due to correlations between ambiguity and frequency, and frequency and length. We generally found large, independent effects, statistically-significant effects of phonotactic probability, length, and frequency. This provides strong evidence that these factors each influence degree of ambiguity, even while controlling for the other factors. This verifies a prediction of the minimal-effort explanation for ambiguity: every factor we tested which we predicted to increase ease of processing, also increased ambiguity.

This is not to say that there is no cost to ambiguity. First, comprehenders must actively use context to disambiguate meaning. However, considerable evidence from language processing indicates that comprehenders are able to quickly use contextual information in the form of discourse context (Altmann & Steedman, 1988; Altmann, Garnham, & Dennis, 1992; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Spivey & Tanenhaus, 1998; Kaiser & Trueswell, 2004; Grodner, Gibson, & Watson, 2005), local linguistic context (Frazier, 1979; Jurafsky, 1996; Gibson, 1998; McDonald & Shillcock, 2003; Frisson, Rayner, & Pickering, 2005; Levy, 2008), or more global world knowledge (Trueswell, Tanenhaus, & Garnsey, 1994; Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003) in disambiguating language. These systems may be just as useful for normal language comprehension, as they are for disambiguating the types of ambiguity discussed in this paper. Comprehenders continually make inferences about what speakers are intending to convey (Grice, 1969; Crain & Steedman, 1985; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Levinson, 2000; Sedivy, 2002), both and what their utterances may mean on literal and pragmatic levels. In fact, Levinson (2000) has argued explicitly that speaker articulation, not hearer inference, is the principal bottleneck in human language. Inference is "cognitively cheap": therefore, normal human communication requires the comprehender to make continual inferences about speaker intention, and does not require the speaker to fully articulate every shade of meaning and resolve every ambiguity.

A more substantial cost for ambiguity arises when inference fails, causing actual confusion. Wasow et al. (2005) list several real world ambiguities causing communicative problems, although they point out they have no way of estimating the frequency with which such situations arise. However, other researchers have claimed that they are vanishingly

rare. Miller (1951) even argued that language only appears ambiguous when we try to examine words out of their normal usage context. Considering the many senses of the word "take," he wrote "Why do people tolerate such ambiguity? The answer is that they don't. There is nothing ambiguous about 'take' as it is used in everyday speech. The ambiguity appears only when we, quite arbitrarily, call isolated words the unit of meaning." Indeed, polysemy and homophony appear to be so well-disambiguated by context that we often consciously notice genuine ambiguity as humorous[4]. Nearly all frequent words have multiple senses but word senses can be disambiguated reasonably well by computational models (Navigli, 2009), even using very simple knowledge of context or heuristics (Gale, Church, & Yarowsky, 1992; Yarowsky, 1993). Similarly, structural ambiguities that slow down human comprehension are extremely rare (Wasow & Arnold, 2003; Jaeger, 2006; Ferreira, 2008; Jaeger, 2010). Therefore, we believe that the potential for miscommunication is rare enough relative to the degree of ambiguity that it is reasonable ignore this communicative cost, at least as an approximation.

Language users do not appear to go to great lengths to avoid linguistic ambiguities, despite actively avoiding conceptual ambiguities. Ferreira, Slevc, and Rogers (2005) found that experimental participants chose to produce descriptions of objects that avoided conceptual ambiguities, such as saying "small bat" rather than just "bat" when a large bat was also present. However, speakers much less often went to similar lengths to avoid purely linguistic ambiguities (such as "baseball bat" when an animal bat was also present). Similarly, when choosing between different syntactic expressions of an intended meaning (such as the whether to omit the "who were" in "the astronauts who were selected...", which would lead to a temporary ambiguity) speakers seem to produce the fuller or clearer expressions no more often or only slightly more often when there is the potential for ambiguity (e.g. Allbritton, McKoon, & Ratcliff, 1996; Arnold, Wasow, Asudeh, & Alrenga, 2004; Ferreira & Dell, 2000; Haywood, Pickering, & Branigan, 2005). These findings suggest that ambiguity is not enough of a problem to real-world communication that speakers would make much effort to avoid it. This may well be because actual language in context provides other information that resolves the ambiguities most of the time. Such information could be prosodic (e.g. Mims & Trueswell, 1999; Snedeker & Trueswell, 2003; Kraljic & Brennan, 2005) (though see Allbritton et al., 1996), or it may be given by context, meaning that in real language use there is rarely much need to actively choose linguistic forms which are unambiguous in isolation (e.g. Ferreira & Dell, 2000; Haywood et al., 2005; Ferreira, 2008).

Our arguments are closely related to *Uniform Information Density* (UID: see Jaeger (2010)), which holds that speakers are more likely to choose words and structures which maintain a roughly constant rate of information transmission. UID and closely related theories have been used to explain phenomena such as discourse-level predictability (Genzel & Charniak, 2002, 2003; Piantadosi & Gibson, 2008; Qian & Jaeger, submitted), syntactic choice (Jaeger, 2006; Levy & Jaeger, 2007; Jaeger, 2010) and reduction (Van Son & Pols, 2003; Aylett & Turk, 2004; Frank & Jaeger, 2008). An ambiguous linguistic form conveys less information about its intended meaning than an unambiguous linguistic form. Therefore, to keep the entropy rate constant, one might choose ambiguous linguistic units which are *less* surprising in other ways which match our findings: UID predicts that am-

[4]Whether or not it is intended as such, as in the headline "Chimpanzee Training Expert to Lecture."

biguous words will be more phonotactically predictable, higher-frequency (less surprising), and shorter (to maintain constant information-per-time). However, we argue that the results presented above are not merely a consequence of UID, though they rely on similar ideas and theoretical basis. Most importantly, UID does not directly predict that efficient language *should* be ambiguous. One could easily design a linguistic communication system in which more informative, or surprising words were shorter, without necessarily making any words ambiguous about, say, their intended sense. Analogously in coding theory, one can construct codes such as Huffman codes for which the length of each code word depends on its log probability, but the code words can map unambiguously to meanings. Additionally, if language is rarely ambiguous in context, it is not clear that UID makes predictions about the out-of-context measures of ambiguity studied here, since UID is a theory of in-context language use. That is, the information rate for an ideal observer does not depend on contextually-unsupported interpretations, since an ideal observer eliminates them. As such, ideal-observer models of UID do not make predictions about ambiguity, but less rational characterizations which consider contextually-unsupported meanings may.

Conclusion

We have provided several kinds of evidence for the view that ambiguity results from a pressure for efficient communication. We argued that any efficient communication system will necessarily be ambiguous when context is informative about meaning. The units of an efficient communication system will not redundantly specify information provided by the context; when examined out of context, these units will appear not to completely disambiguate meaning. We have also argued that ambiguity allows efficient linguistic units to be preferentially re-used, decreasing the overall effort needed to use a linguistic system.

We tested predictions of this theory by showing that ambiguity allows re-use of the easiest linguistic units. These results are hard to explain with anything other than a theory based on efficient communication: what theory would posit that ambiguity should preferentially be found in these linguistic units, but not that it results from pressure for efficiency? Our results argue for a rational explanation of ambiguity and demonstrate that ambiguity is not mysterious when language is considered as a cognitive system designed in part for communication.

References

Allbritton, D., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 714–735.

Altmann, G., Garnham, A., & Dennis, Y. (1992). Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, *31*(5), 685–712.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing* 1. *Cognition*, *30*(3), 191–238.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. Routledge.

Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, *23*(4), 495–527.

Arnold, J., Wasow, T., Asudeh, A., & Alrenga, P. (2004). Avoiding attachment ambiguities: The role of constituent ordering* 1. *Journal of Memory and Language*, *51*(1), 55–70.

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, H. R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database. release 2 (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania.

Chomsky, N. (2002). An interview on minimalism. *N. Chomsky, On Nature and Language*, 92–161.

Cover, T., & Thomas, J. (2006). *Elements of information theory.* John Wiley and sons.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. *Natural language parsing*, 320–358.

Demberg, V. (2010). *A broad-coverage model of prediction in human sentence processing.* Unpublished doctoral dissertation, University of Edinburgh.

Fellbaum, C., et al. (1998). *WordNet: An electronic lexical database.* MIT press Cambridge, MA.

Ferreira, V. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation*, *49*, 209–246.

Ferreira, V., & Dell, G. (2000). Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production* 1. *Cognitive Psychology*, *40*(4), 296–340.

Ferreira, V., Slevc, L., & Rogers, E. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, *96*(3), 263–284.

Ferrer i Cancho, R. (2006). When language breaks into pieces: A conflict between communication through isolated signals and language. *Biosystems*, *84*(3), 242–253.

Ferrer i Cancho, R., & Loreto, V. (in prep). Conditions for the emergence of combinatorial communication.

Ferrer i Cancho, R., & Solé, R. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 788.

Frank, A., & Jaeger, T. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the cognitive science society.*

Frazier, L. (1979). On comprehending sentences: Syntactic parsing strategies. *ETD Collection for University of Connecticut*.

Frisch, S., Broe, M., & Pierrehumbert, J. (n.d.). *Similarity and phonotactics in arabic.*

Frisson, S., Rayner, K., & Pickering, M. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Learning, Memory*, *31*(5), 862–877.

Gale, W., Church, K., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the workshop on speech and natural language* (p. 237).

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Hierarchical/Multilevel Models.* Cambridge: Cambridge University Press.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 199–206).

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of empirical methods in natural language processing* (pp. 65–72).

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Givón, T. (2009). *The genesis of syntactic complexity: diachrony, ontogeny, neuro-cognition, evolution.* John Benjamins Publishing Co.

Graff, P., & Jaeger, T. (2009). Locality and feature specificity in ocp effects: Evidence from aymara, dutch, and javanese. In *Proceedings of the main session of the 45th meeting of the chicago linguistic society, chicago, il.*

Grice, H. (1969). Utterer's meaning and intention. *The philosophical review*, *78*(2), 147–177.

Grodner, D., Gibson, E., & Watson, D. (2005). The influence of contextual contrast on syntactic processing: evidence for strong-interaction in sentence comprehension. *Cognition*, *95*(3), 275–296.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*(2), 274–307.

Haywood, S., Pickering, M., & Branigan, H. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science*, *16*(5), 362.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Jaeger, T. (2006). *Redundancy and syntactic reduction in spontaneous speech.* Unpublished doctoral dissertation, Stanford University.

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

Juba, B., Kalai, A., Khanna, S., & Sudan, M. (2011). Compression Without a Common Prior: An Information-theoretic Justification for Ambiguity in Language. In *2nd symposium on innovations in computer science.*

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Kaiser, E., & Trueswell, J. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, *94*(2), 113–147.

Kamide, Y., Altmann, G., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kraljic, T., & Brennan, S. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, *50*(2), 194–231.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the twentieth annual conference on neural information processing systems.*

MacKay, D. (2003). *Information theory, inference, and learning algorithms.* Cambridge University Press.

Manin, D. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, *32*(7), 1075–1098.

Marslen-Wilson, W., Levy, E., & Tyler, L. (1982). Producing interpretable discourse: The establishment and maintenance of reference. *Speech, place, and action*, 339–378.

McCarthy, J. (1985). *Formal problems in semitic phonology and morphology.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

McCarthy, J. (1986). Ocp effects: gemination and antigemination. *Linguistic inquiry*, *17*(2), 207–263.

McDonald, S., & Shillcock, R. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648.

Miller, G. (1951). *Language and Communication.* McGraw-Hill.

Mims, K., & Trueswell, J. (1999). Do speakers help listeners? Prosodic cues, lexical cues, and ambiguity avoidance. In *Poster presented at the twelfth annual cuny conference on human sentence processing, new york.*

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*(6), 1191–1253.

Piantadosi, S., & Gibson, E. (2008). Uniform Information Density in Discourse. In *Talk presented at the twenty-first annual cuny conference on human sentence processing.*

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and brain sciences*, *13*, 707–784.

Qian, T., & Jaeger, T. (submitted). Entropy profiles in language: A cross-linguistics investigation. *Entropy*.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*, 372–422.

Reder, L., Anderson, J., & Bjork, R. (1974). A semantic interpretation of encoding specificity. *Journal of Experimental Psychology*, *102*(4), 648–656.

Roland, D., Elman, J., & Ferreira, V. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*, *98*(3), 245–272.

Sedivy, J. (2002). Invoking discourse-based contrast sets and resolving syntactic ambiguities. *Journal of Memory and Language*, *46*(2), 341–370.

Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 623–656.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130.

Spivey, M., & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology, Learning, Memory and Cognition*, *24*, 1521–1543.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632.

Tily, H., & Piantadosi, S. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference. amsterdam, the netherlands.*

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Memory and Language*, *33*, 285–318.

Van Son, R., & Pols, L. (2003). How efficient is speech? *Proc. of the Institute of Phonetic Sciences*, *25*, 171–184.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*(1), 79–112.

Wasow, T., & Arnold, J. (2003). Post-verbal constituent ordering in English. *Determinants of grammatical variation in English*, 119–154.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. *Morphology and The Web of Grammar: Essays in Memory of Steven G. Lapointe. CSLI Publications*.

Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the workshop on human language technology* (p. 271).

Zipf, G. (1936). *The psychobiology of language*. London: Routledge.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.