Language and Linguistics Compass 10/9 (2016): 397–412, 10.1111/lnc3.12196

Information-theoretical Complexity Metrics

John Hale*
Department of Linguistics, Cornell University

Abstract

Information-theoretical complexity metrics are auxiliary hypotheses that link theories of parsing and grammar to potentially observable measurements such as reading times and neural signals. This review article considers two such metrics, Surprisal and Entropy Reduction, which are respectively built upon the two most natural notions of 'information value' for an observed event (Blachman 1968). This review sketches their conceptual background and touches on their relationship to other theories in cognitive science. It characterizes them as 'lenses' through which theorists 'see' the information-processing consequences of linguistic grammars. While these metrics are not themselves parsing algorithms, the review identifies candidate mechanisms that have been proposed for both of them.

1. A Mathematical Theory

Information theory, in the sense of Shannon (1948), is a mathematical theory. This means that basic concepts such as 'message' are not further defined but rather may receive a variety of interpretations: a sequence of dots and dashes as in Morse code or a sequence of light intensities as in black and white television, to take just two examples from Shannon's paper. This mathematical theory, originating in telecommunications, has turned out to be widely applicable across many different fields. In the language sciences, including linguistics, it has been enjoying a revival since the 1990s. This revival makes it clear that information theory indeed applies quite generally to questions of language structure, acquisition and processing. Malouf (forthcoming) provides an extensive bibliography. This review focuses more narrowly on COMPLEXITY METRICS in language comprehension. These are linking hypotheses that relate theories of syntactic parsing to empirical data such as reading times or blood oxygen levels in the brain.[1] For many researchers, they are appealing precisely because they suggest a way of understanding the particularities of human language processing in terms of a general mathematical theory.

2. Complexity Metrics and Incrementality

Generally speaking, a complexity metric is something that quantifies how difficult it is to perceive a linguistic expression. This characterization includes any rule that relates theorized parsing mechanisms to word-by-word measures of comprehension difficulty.[2] Kaplan's (1972) 'number of transitions made or attempted' (page 89) is a classic example. But there are many others: the number of unresolved case-roles at a given point (Morrill 2000; Stabler 1994), the number of phrase structure nodes in memory (Frazier 1985; Yngve 1960), or the latency of a memory retrieval operation motivated by a syntactic dependency (Lewis and Vasishth 2005). The length of this catalog testifies to the important role that complexity metrics have always played in relating theorized mechanisms to observable data. One commonality among the metrics just cited is the fact that they make per-word, rather than per-sentence predictions. It is this 'incremental' feature that is

important for online comprehension studies. By contrast, the ordered list of construction types given in Caplan, Baker, and Dehaut (1985: 123) or the ranking of larger and smaller domains suggested in Hawkins (2010) both offer their predictions at the level of the sentence, rather than the word. They are complexity metrics but not incremental ones.

3. Deriving Predictions About Potential Observations

Surprisal and Entropy Reduction are incremental complexity metrics that predict how difficult each word should be as it is perceived in time. They are information-theoretical insofar as they view sentences as random events. From this technical perspective, words are symbols that appear with a certain probability. The revelation of each new word-symbol as being the particular symbol that it is constitutes a sub-event with a quantifiable information value. Both metrics suppose that greater information value should relate to greater processing difficulty. However they differ in the precise formulation of 'information value' that they apply. Both can be thought of as summaries of the transition between a word and its successor, but they are mathematically different and can derive contrasting empirical consequences e.g. on relative clauses (see Section 5.4). In typical psycholinguistic modeling practice, one computes the value of the metric at each word-position in a stimulus sentence. These predictions are then compared with observed measures of comprehension effort, such as reading times, scalp potentials, or blood oxygen levels in particular brain areas. If the per-word information values match up well with the observed measures then we say that the observed data support the conjunction of the complexity metric and whatever defined the probabilities in the first place.

Of course the key issue is what defines the probabilities. This boils down to the question of language model, to which we now turn.

4. Language Models

Information-theoretical complexity metrics like Surprisal and Entropy Reduction are defined in terms of probability distributions having to do with the transition from one word to the next. The specification of these distributions is called a language model. The word 'language' in this name alludes to the classic conceptualization of a language as a set of strings of words (e.g. Chomsky 1956). In this narrow sense, a language model is simply an assignment of probability to each string in this set. The question is, how to define these probabilities?

4.1. The Obvious way is Vulnerable to a Famous Critique

Perhaps the most straightforward way to assign probabilities to strings is to probabilistically generate each word, one after the other. Under this arrangement, the probability of a successor word is defined in terms of the previous words that have already been generated. The general view is that of a table keeping track of (a) the last *n-1* words, (b) the n^{th} word and (c) its conditional probability, $P(w_i \mid w_{i-1}\ w_{i-2}\ \cdots\ w_{i-(n-1)})$. Rows of this table, namely combinations of possible successor words sharing the same left-context, serve to define a discrete distribution. One can then ask how surprising is a specific word w_i, how uncertain is the distribution as a whole, et cetera.

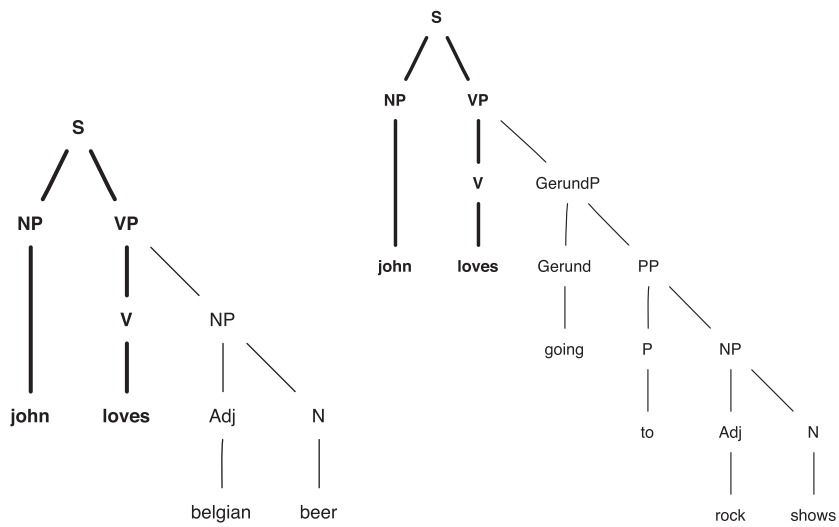
This view is appealingly straightforward, but as a scientific proposal, it leaves much to be desired. In fact, it is exactly the Markov model of language that Chomsky (1956) criticized. The core of Chomsky's critique is that, in real human language, words

influence each other at arbitrary distances in a way that Markov models cannot properly capture. In cases of grammatical agreement, coordination or relative clause formation, for instance, the presence or absence of an upcoming word *depends* upon words that may be quite far back in the stream. If the dependency is not fulfilled, the probability should be essentially zero. For dependencies wider than n, it will be impossible for an n^{th}-order Markov model to assign realistic probabilities.[3] The problem is not a matter of degree. Rather, it is the Markov property itself that fails to be fulfilled by natural language.

4.2. The Conditional Probability of a Grammatical Derivation

To overcome this problem, speech recognition pioneers like Jelinek and Lafferty (1991) turned to probabilistic grammars. In a probabilistic grammar,[4] conditional probabilities are associated with rewriting rules, so that derivation is now a branching process as shown in Figure 1. This Figure shows two derivations which share a common initial substring, **john loves**. In one derivation, the symbol VP is rewritten by a rule VP → V NP, whereas in the other, VP is rewritten by a different rule VP → V GerundP. Estimating the probabilities associated with these rule alternatives can be done, for instance, using parsed corpora as in Hale (2001) and Levy (2008a).

Viewing derivation as a branching process in this way means that derivation trees are values of a random variable, X. Conditioning on a particular initial substring, such as **john loves**, isolates a particular subset of these derivations. This is analogous to selecting rows in a conditional probability table, except that the subset may itself be infinite. A probabilistic grammar defines a distribution on this subset just as it does for the entire language. It is this distribution, or rather, before/after pairs of distributions on either side of the latest



(a) Derivation d_1 (b) Derivation d_2

Fig. 1. Two derivations that share the same initial substring. The probability of an initial substring is the sum of the probabilities of all derivations of that substring on a given probabilistic grammar, e.g. P(**john loves**) is defined to be $P(d_1) + P(d_2) + \cdots$ for all derivations d_i whose yield begins with **john loves**.

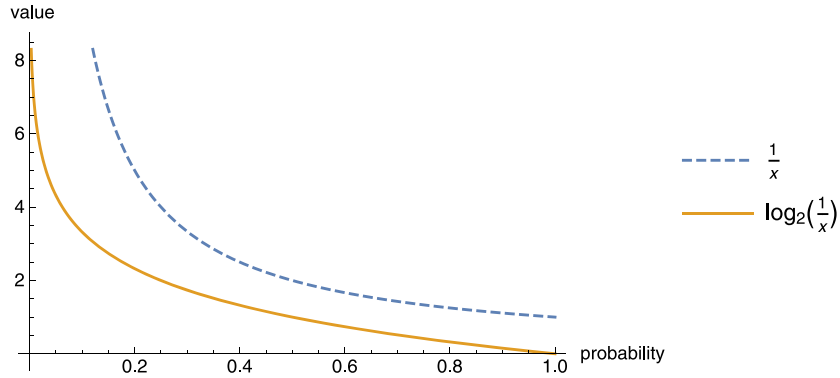


Fig. 2. Reciprocal and surprisal in the interval (0, 1].

word, that information-theoretical metrics summarize. The next section applies this perspective to Surprisal, before moving on to Entropy Reduction in Section 6.

5. Surprisal

5.1. Definition

The term 'surprisal' dates back to Tribus (1961), who used it to talk about the logarithm of the reciprocal of a probability. It is one way to characterize the information value of an observed event (Section 6 presents another). For a generic random variable Y, the surprisal of an outcome $Y = y$ is

$$\log_2\left(\frac{1}{P(y)}\right) \tag{1}$$

Note that the argument to the logarithm in this expression is the reciprocal of a probability. Since probabilities are between 0 and 1, the smallest value of this reciprocal corresponds to the largest probability i.e. $\frac{1}{1}$. As probabilities get closer to zero, their reciprocal gets larger and larger. Figure 2 shows this pattern using a dashed line. Taking the logarithm of these reciprocals pulls the curve down a bit but retains the idea of assigning higher values to lower probabilities.[5] In a nutshell, surprisal is higher for low-probability events.[6]

To apply this formal definition, one needs to fix upon a particular event Y: the appearance of a word as the next symbol in a string. This implies the existence of two strings, both anchored at the beginning of a sentence, but where one string is exactly one symbol longer than the other. These are 'initial substrings' of the sort indicated in **bold** in Figure 1. Both of them may be associated with conditional probability distributions using a grammar. As suggested in that Figure, the support of these distributions is precisely the set of derivations that derive them. The total probability mass that they assign is known as the 'prefix probability,' and it is the ratio of these prefix probabilities that gives the transition probability of the next symbol. This ratio defines $P(y)$ in the surprisal complexity metric. This is written out below in Equation 2 where these prefix probabilities are cartooned as sums ($\sum$) either before the successor word or after it.

$$\log_2\left(\frac{1}{P(y)}\right) \quad \text{let } y \text{ be the ratio of prefix probabilities} \quad (2)$$
$$= \log_2\left(\frac{1}{\frac{\sum \text{after}}{\sum \text{before}}}\right)$$
$$= \log_2\left(\frac{\sum \text{before}}{\sum \text{after}}\right)$$
$$= -\log_2\left(\frac{\sum \text{after}}{\sum \text{before}}\right)$$

To illustrate this application of the generic surprisal idea in Equation 1 to grammatical derivations, as in Equation 2, Figure 3 shows a pair of hypothetical probability distributions of the sort that might be generated by a probabilistic grammar. Each bar corresponds to a derivation, such as those in Figure 1. The height of the bars represents each derivation's probability. The histogram on the left depicts the distribution at the shorter initial substring. The arrow at the top symbolizes the transition from one word to the next in the course of incremental parsing. With the appearance of the next word, the set of available derivations contracts. Some derivations are incompatible with the new word, as shown in the histogram on the right which corresponds to the longer substring. Successive transitions zero-out derivations in this way until, in an unambiguous sentence, presumably only one is left. Each time this happens, the probability assigned to the missing bars is lost.[7] With surprisal, the total amount lost corresponds logarithmically to the predicted comprehension difficulty.

5.2. Empirical Support for Surprisal

Surprisal has seen broad success across many different methodologies. In eye-tracking, several different parsing mechanisms and grammar types converge on the idea that people read more slowly on words whose surprisal value is higher (Boston et al. 2008; Demberg and Keller 2008; Rauzy and Blache 2012, inter alia). Scanpaths recorded this way are more irregular

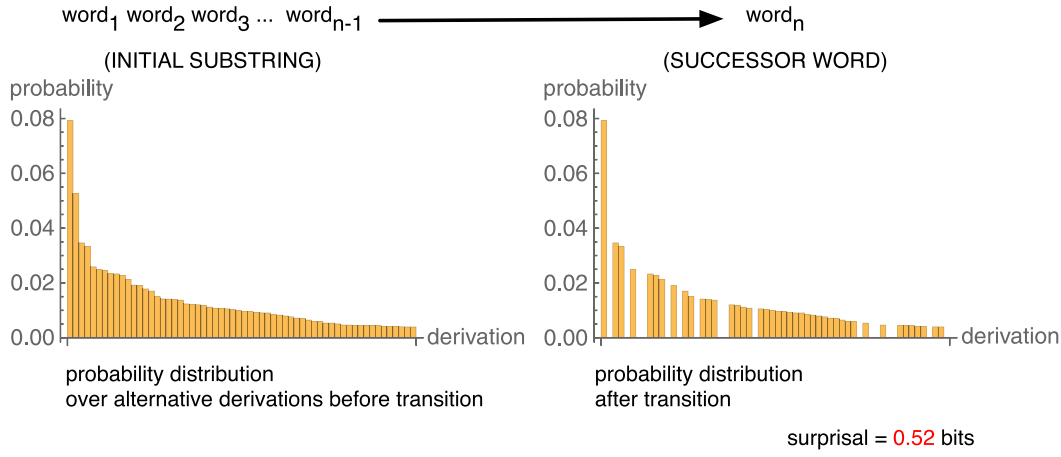


Fig. 3. Transitioning from word to word, derivations are ruled out. The probability of this transition itself can be defined as a ratio of sums, $\frac{after}{before}$. The log of the reciprocal of this ratio is the surprisal of the next word. The greater the probability mass ruled out, the higher the surprisal.

when surprisal is higher (von der Malsburg, Kliegl, and Vasishth 2015). With event related potentials, similarly, surprisal is positively related to the amplitude of the N400 component (Frank et al. 2015). And in functional MRI, surprisal from phrase-structured language models predicts the timecourse of activation in several different brain areas including anterior temporal lobe (Brennan et al. 2016; Hale et al. 2015; Henderson et al. 2016). Timecourses from several other brain areas appear to correlate well with surprisal values from *n*-gram models (Willems et al. 2015).

Surprisal is successful empirically because it accounts naturally for frequency effects. Such effects have long been recognized in behavioral studies. To take just one example, Thibadeau, Just, and Carpenter (1982) found unigram probability to be a useful predictor of eye-fixation duration. Surprisal generalizes this idea from single words to larger domains, e.g. where $n > 1$. It extends the idea that 'rare-implies-more-difficult' to syntactic phrases. However, there's more than one way to do it! In making this extension, several design decisions present themselves. Many of these reflect classic issues in cognitive science for which the complexity metric itself offers no easy answers. The following subsections take up two such issues, focusing primarily on their implications for Surprisal, although many of the same considerations apply equally to Entropy Reduction (see Section 6).

5.3. Crosscutting Design Decisions

5.3.1. Lexicalization

Linguistics has wrestled since the 1970s with lexicalism, the idea that grammatical analyses should be sensitive to the idiosyncrasies of words themselves. The decision to be lexicalist or not carries over directly into information-theoretical complexity metrics. For instance, the phrase-structured language model of Section 4.2 is partially lexical. This means that some probabilities are contributed by preterminal rules having to do with real words e.g. $\mathrm{N} \rightarrow \mathsf{beer}$ or $\mathrm{V} \rightarrow \mathsf{loves}$, while others come from phrasal rules like $\mathrm{NP} \rightarrow \mathrm{Adj\ N}$ or $\mathrm{PP} \rightarrow \mathrm{P\ NP}$ that refer only to syntactic categories. Both types of rules become associated with numerical weights in the course of fitting a probabilistic grammar (see references in endnote 4). It is also possible, however, to fit the numerical parameters of a parser or grammar to part-of-speech tag sequences. This suppresses the lexical contribution, allowing only the phrasal rules to contribute to the metric.

Rather than suppressing lexicalism, the other possibility is to embrace it. One can lexicalize more and more rules, including those classically viewed as independent of particular words (see the example in Hale 2014, Chapter 2). Such an approach holds out hope of accounting for detailed lexical effects in human sentence processing (as suggested by Ford, Bresnan, and Kaplan 1982; MacDonald 1994). However, it does so by entangling the notion of *predictability* with that of *rare words*. Predictability is presumably what is measured in the Cloze task, where participants must guess successor words given an initial substring as a prompt (Taylor 1953). On the other hand, lexical frequency or more precisely unigram probability is by definition based on individual occurrences in a corpus. These two factors seem to be psychologically separable and might be better understood in isolation from each other (Staub 2015). Another downside is the greater number of numerical parameters in lexicalized grammars. Allowing for more interactions between phrases and words makes them more difficult to estimate in practice.

5.3.2. Parallel vs. Serial Processing

Another key design decision revolves around approximations of the prefix probability. In the Hale (2001) formulation, all structural alternatives affect the metric. One might interpret this in terms of a fully parallel parsing mechanism that considers all possible analysis paths. Under this regime, an analysis remains in play until it is definitively ruled out by some observed word. This is exactly what happens in garden path sentences, where the 'pain' of committing to the globally correct interpretation is deferred until the disambiguation point. In this way, surprisal can emulate theories of attachment preferences, such as Frazier and Fodor's Garden Path Theory.[8]

An alternative is to impose some parallelism limit. Parsing accuracy comparisons suggest that even a restriction to 3 or 4 syntactic analyses leads to good performance on the Wall Street Journal (Brants and Crocker 2000). In light of this, Boston et al. (2011) parametrically varied the amount of parallelism available to a particular parser. Eyetracking predictions derived from this parser via surprisal got better and better as the parser considered more alternative pathways.[9] This convergence toward an ideal, where memory is unlimited, suggests that the metric itself is on the right track.[10]

5.4. Empirical Challenge: Relative Clauses

On the other hand, there is data suggesting that surprisal's simple equation between probability and difficulty may actually be too simple. The empirical challenge comes from relative clauses (RCs). In a relative clause, there is a missing element, marked *t* in Example 3 below.

a. The reporter $[_{\text{RC}}$ that *t* attacked the senator$]$ admitted the error
b. The reporter $[_{\text{RC}}$ that the senator attacked *t*$]$ admitted the error (3)

In transformational grammar, one would say that the missing element has been 'relativized' leaving a 'trace'. In a particular language, relativization applies to a delimited subset of grammatical relations: subject, direct object, indirect object et cetera. These subsets, to which relativization may apply, are organized into a scale such that processing difficulty varies inversely with typological ubiquity in the world's languages (Keenan and Hawkins 1987). Within psycholinguistics, attention has focused on the first two points of this scale, subject- and object- extracted RCs. These are shown in Example 3 using stimuli from King and Just (1991).

Object-extracted RCs as in 3b are rare in natural language corpora, and as a result probabilistic grammars fitted to them end up including a low-probability rule. This is crucial, because as Figure 2 suggests, lower probability implies higher surprisal. Such a rule would be used in the derivation of object-extracted RCs only. As a consequence of this difference, surprisal correctly predicts that 3b should be the more difficult of the two constructions. But as an INCREMENTAL complexity metric, it is less accurate: the vanilla version of surprisal would predict effort at the point where the low-probability rule becomes obligatory. This happens at the left-edge of the relative clause (notated $[_{\text{RC}}$ in Example 3) rather than at the embedded verb *attacked*. This is where Grodner and Gibson (2005) report a reading time slowdown. While Staub (2010) later found effects at this leftmost position, the general feeling during the early 2000s was that surprisal had failed to derive an important part of

the data pattern. Expressing this general feeling, Levy (2008a): page 1166) refers to 'mixed results'.

> One way of interpreting these mixed results is to hypothesize that surprisal has a major effect on word-by-word processing difficulty, but that truly non-local (i.e., long-distance) syntactic dependencies such as relativization and WH-question formation are handled fundamentally differently […]

The suggestion is to back off to a more complex, two-factor model where surprisal's role is somewhat curtailed.[11]

It is of course possible that some other grammar type or parsing mechanism would yield better surprisal predictions, by specifying a different order in which the relevant probabilities take effect. But investigation along this line faltered under the assumption that the embedded verb is the primary locus of processing difficulty. This mismatch between theory and data motivated the search for another complexity metric, one that would provide a one-factor explanation of the difficulty profile in relative clauses.

6. Entropy Reduction

6.1. Definition

Entropy reduction is that metric. Unlike surprisal's logarithmic transformation of probability, it instead formalizes the information value of a word using the notion of ENTROPY. This quantity is the centerpiece of information theory. Shannon (1948) suggests its name by analogy to thermodynamics.

> The quantity which uniquely meets the natural requirements that one sets up for 'information' turns out to be exactly that which is known in thermodynamics as *entropy*.

The entropy of a random variable X is defined below in Equation 4.

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x) \tag{4}$$

The entropy H quantifies uncertainty in X's probability distribution. For instance, a 100-sided die has greater uncertainty than a 6-sided die. When outcomes are equiprobable, entropy is at a maximum. When alternatives are unequally weighted, it is easier to guess the outcome; we become less uncertain. The core intuition of the Entropy Reduction complexity metric is that this sort of 'information gain' should index observable human comprehension difficulty.

Entropy Reduction was inspired by and can be viewed as a generalization of Den and Inoue's (1997) Verb Predictability Hypothesis. Whereas Den and Inoue were concerned with the size of the garden path effect at a sentence-final verb, Entropy Reduction applies to all positions and all categories. Taking X again to be derivations on a probabilistic grammar, we ask: by how much does knowledge of the initial substring $Y = y$ reduce uncertainty? The answer is the information value I:

$$I(X; y) = H(X) - H(X|y) \tag{5}$$

In Equation 5, the quantity $H(X|y)$ is simply the entropy of the subset of derivations sharing an initial substring. Analogous to the way prefix probabilities were divided in Section 5.1, subtractions of these conditional entropies $H(X|y)$ define the informational contributions of

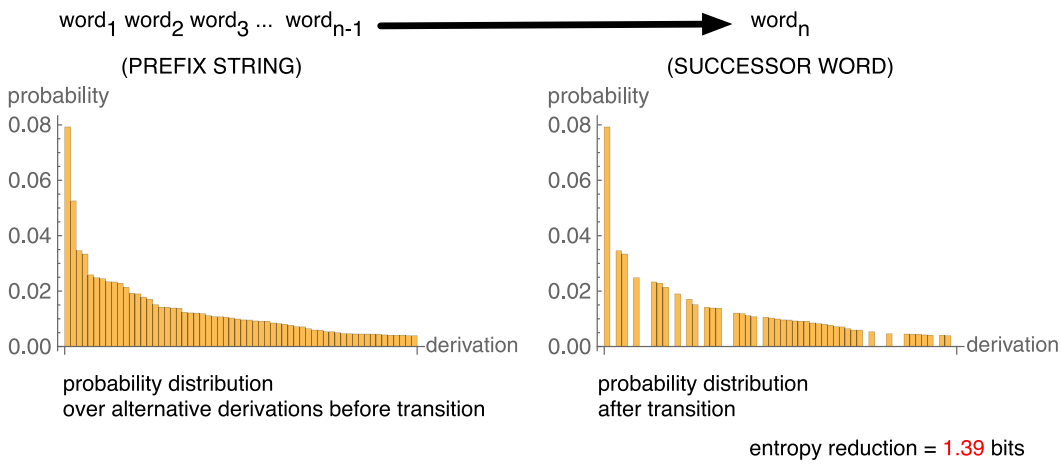


Fig. 4. The same 'before' and 'after' distributions as in Figure 3 yield different information values. Entropy reduction is the downward change, if any, between *H*(*before*) and *H*(*after*), whereas surprisal is the log-ratio of sums of these distributions.

particular words. If the conditional entropy goes down from one word to the next, then grammatical uncertainty has been reduced. The interpretation is that the comprehender has done information-processing work. On the other hand, it is also possible for entropy to go up. This happens when the next word opens up possibilities that are more uncertain than on average. In such cases, the comprehender has done no work: no progress has been made towards the goal of a unique reading.[12]

Figure 4 repeats Figure 3 in order to emphasize the point that surprisal and entropy reduction yield different information values, even from the same grammatical probability distributions. The two metrics summarize the transition between distributions in different ways, leading to different predictions about incremental processing difficulty. Appendix A provides a worked example showing how contrasting predictions follow from the same grammar. These specific cases exemplify the general point, developed in Blachman (1968), that entropy reduction and surprisal are different conceptions of information value.

6. 2. Empirical Support

Entropy reduction derives the comprehension difficulty profile across a wide range of constructions, including relative clauses (Chen and Hale under review; Hale 2003; Hale 2006; Yun et al. 2015). It has been applied with naturalistic text stimuli (Frank 2013; Wu et al. 2010) as well as with controlled experimental materials (Linzen and Jaeger 2015). beim Graben and colleagues (2008, 2000) show how ERP components such as the N400 and P600 can be understood as entropy reductions in an underlying dynamical system.

Using techniques due to Bar-Hillel, Perles, and Shamir (1964) and Nederhof and Satta (2008), entropy reduction can be computed for a wide array of formal grammars. Freely available software (Chen et al. 2014) facilitates this computation, even with expressive grammars that directly define syntactic movement.

For instance, Yun et al. (2015) model the difficulty profile of Chinese, Japanese and Korean relative clauses using a grammar that directly represents movement. In these languages with prenominal relatives, filler/gap dependencies such as the ones between *t* and *reporter* in Example 3 are arranged in such a way that their processing asymmetry cannot be explained by simple memory-based approaches. Yun et al. (2015, §2) go into considerably greater depth

on this point, with summaries of the relevant experimental studies and a taxonomy of available theories. The grammatical approach that these authors propose combines Minimalist syntactic analyses with the Entropy Reduction complexity metric to derive word-by-word predictions across all three languages. The pattern in all cases accords with the observation of a Subject Advantage in RC processing, regardless of whether the RC is prenominal or postnominal. The explanation depends on language-particular facts, for instance regarding the availability of argument omission, the tendency of verbs to be transitive or intransitive and the likelihood of optional modifiers. It echoes Hale (2003) where the explanation of the corresponding English RCs turned on the possibility of recursive modification e.g. by postnominal RCs. The common element across this account of prenominal and postnominal RCs is the complexity metric itself.

On a conceptual level, entropy reduction takes one step away from the externalism of surprisal by focusing on an internal aspect of parser states, namely their uncertainty. From this point of view, the distribution over alternative analyses matters in a way that it does not for surprisal. This research program is naturally extended by adding more information about an agent's goals, such that these internal probabilities align with expected payoffs, becoming 'utilities' (Calvillo and Crocker 2015; Hale 2011; Lewis, Shvartsman, and Singh 2013).

7. Mechanisms

It is foundational, in cognitive science, to differentiate between levels of explanation. Marr (1982) distinguishes between a higher 'computational' level and a lower 'algorithmic' level. These two levels are simultaneous co-descriptions of the same organism, at greater and lesser degrees of abstraction. Theories that define what computation the organism is actually doing are stated at the higher level, while theories of how that computation is effected occupy the lower level.[13]

Complexity metrics like Entropy Reduction and Surprisal, then, are computational-level theories: they specify what the difficulty level of parsing a word should be in terms of structures defined in a language model. In order to offer a psychological PROCESS MODEL, one must characterize a mechanism that operates in accordance with these metrics — a compatible co-description.

Hale (2014, Chapter 8) advances a mechanistic interpretation of surprisal, viewing it as a consequence of the Chunking Theory of Learning (CTL) (Rosenbloom and Newell 1987). The CTL supposes that cognitive operators can fuse together in the course of practice. In this way, what used to require multiple steps is now accomplished by a macro-operator that applies all at once, in one step.[14] Analyzing the CTL, Rosenbloom and Newell show how it necessarily derives power-law practice curves, in environments where the probability of a 'pattern' decreases as the pattern size increases. This is certainly true in natural language. They remark (page 227) upon the ease with which a power law can be mimicked by a purely logarithmic function. Of course, this is exactly what surprisal is (Equation 1).

The suggestion is that what we observe as surprisal 'effects' in language could come about through the operation of a highly general chunking mechanism, the same mechanism that explains practice in other cognitive domains. Support for this suggestion comes from analyses of eyetracking data in Hale (2014). These analyses considered triplets of parser actions, aligned in time to particular words in English and French newspaper text. An independent estimate of the degree to which these triplets 'cohere', i.e. were likely to be chunked by a chunking mechanism, was a positive predictor of eye-fixation duration. The interpretation is that a person reads familiar sentence structures faster because the cognitive operations that build those

structures have been folded into macro-operators. From this perspective, construction grammar would be a correct description of the chunked representations at the same time as generative grammar remains a correct description of the un-chunked representations. Further work along this line might consider alternative training regimes that experimentally induce chunking.

For entropy reduction as well, candidate mechanisms have been proposed. Hale (2011, §5) offers an 'entropophobic' search heuristic that tries not to get stuck in syntactic categories that could be very complex. When this heuristic guides a phrase structure parser, transition counts (see endnote 2) line up with the observed behavioral contrast between two well-known types of garden-path effects, one relatively mild and the other more severe. An alternative mechanism would be a neural net whose intermediate states define a probability distribution as in Frank (2013).

8. Past and Future

The revival of information-theoretical complexity metrics is a revival of ideas that fascinated information-processing psychologists like Hick (1952), Attneave (1959) and Garner (1962), as well as pre-generative linguists like Charles F. Hockett (1953). In the 1950s, information theory seemed to offer limitless new vistas for understanding human language. But as Luce (2003) details, it fell out of favor. Among other factors, Luce places the blame on a lack of structure in the models that were at that time under consideration:

> The elements of choice in information theory are absolutely neutral and lack any internal structure. That is fine for a communication engineer….[but] by and large, however, the stimuli of psychological experiments are to some degree structured, and so, in a fundamental way, they are not in any sense interchangeable.

In the 21st century, we now know how to build structured probabilistic models. Grammars are just one example of a probabilistic model that is at the same time sensitive to the statistics of the environment and also able to generalize in a categorical, rule-governed way. Now, unlike in the 1950s, we can assign both an information-theoretic interpretation (e.g. the choice of this symbol is worth γ bits) as well as a detailed linguistic interpretation (this rule introduces a relative clause) to the same mathematical entity. By combining contributions from disciplines that are traditionally separate, we advance cognitive science.

Acknowledgement

This research was supported by NSF CAREER award 0741666.

Notes

* Correspondence address: John Hale, Department of Linguistics, Cornell University. E-mail: jthale@cornell.edu

[1] Delimiting the focus in this way, I nonetheless encourage readers to approach information theory on its own terms; John R. Pierce (1980) provides a nontechnical book-length introduction. Kornai (2007, §7.2) is a shorter, more mathematical treatment embedded in a broader discussion of linguistic complexity. Readers interested in phonological applications should consult Goldsmith (2000) or Hume and Mailhot (2013). For morphology, see Milin et al. (2009). Information theory also figures prominently in models of sentence misperception (Gibson, Bergen, and Piantadosi 2013; Levy 2008b),

language production (Jaeger 2010; Keller 2004) and language learning (Chater et al. 2015). This modeling literature is sometimes polemical, arguing for a 'rational analysis' of human psychology founded on a Bayesian interpretation of probability. In what follows, I set aside this polemic since information theory is compatible with any philosophy of probability.

[2] Kaplan (1972) offers an explicit parsing model called an Augmented Transition Net (ATN), which transitions from state to state as it makes its way through a sentence. This model associates one additional unit of perceptual effort with each transition. This claim is particularly interesting given that the transitions are *ordered* and some degree of backtracking is invariably necessary in the search for a successful analysis.

[3] Partee et al. (1993, §17.3.2) present a modernized version of the argument against Markov models.

[4] For a gentle introduction to probabilistic grammars see Chapter 7 of Hale (2014), Chapter 14 of Jurafsky and Martin (2008) or Chapter 3 of Levelt (1974).

[5] The choice of a logarithmic measure is defended in the Introduction to Shannon's Mathematical Theory of Communication (1948). Specifying that the base of the logarithm is 2 is a calibration which means that the surprisals are counted in bits, just like computer memory. Through linear regression one can fit these information theoretical predictions in bits to particular dependent variables, e.g. reading times in milliseconds.

[6] This logarithmic difficulty rule is analogous to the Hick-Hyman law, which characterizes the time a person takes to decide between multiple choices. See Pierce (1980: 230).

[7] The distributions illustrated in Figure 3 have not been renormalized; in this sense, the picture is drawn from the point of view of the grammar which assigns probabilities to full derivations.

[8] Frazier and Clifton (1996, Chapter 1) provide a concise summary of Garden Path Theory. For the standard formalization due to Pereira and Shieber, see Hale (2014, Chapter 4).

[9] The surprisal analysis in Boston et al. (2011) takes as data five different eyetracking measures from the Potsdam Sentence Corpus (Engbert et al. 2005).

[10] Bresnan (1982) exhorts cognitive scientists to 'discover ways of showing that the actual behavior of real native speakers converges on the ideal behavior predicted by our grammatical theory, as interfering performance factors are reduced' (page xxiii). Kaplan later explained this further, saying:

> The basic idea is that you can evaluate theories of grammar-based processing as to whether their behavior corresponds to the behavior of an ideal native speaker in the limit as the amount of available processing resources goes to infinity. Of course, the behavior of an ideal native speaker, one who knows his language perfectly and is not affected by restrictions of memory or processing time, lapses of attention, and so forth, is difficult to observe. But as psycholinguistic methods and technologies improve, we can imagine doing experiments in which we somehow vary the cognitive resources of real speakers and hearers, by removing distractions, giving them scratch-pad memories, etc. We can then take the limiting, asymptotic behavior of real speakers as approximations to the behavior of the ideal. A grammar-based processing model which, when given more and more computational resources, more and more accurately simulates the behavior of the ideal has the 'ideal-convergent' property (Kaplan 1995: page 344).

[11] Demberg and Keller (2009) go on to propose just this sort of two-factor model.

[12] An approximation of entropy reduction restricts consideration to just uncertainty about the next word, rather than the whole derivation. Here again, a theorist must decide how to structure the left context: either by phrase, as in Roark et al. (2009), or by word, as in Willems et al. (2015).

[13] The difference between Marr's uppermost two levels clarifies the respective roles of generative grammar and psycholinguistics in one classic conception of cognitive science. See Hale (2014, Chapter 1) for a discussion of Marr's framework as applied to language.

[14] Such operator-chunking in fact contributed to NL-Soar's speed, bringing it into the time band of human language comprehension (see Lehman, Lewis, and Newell 1991 and Lewis 1993, Chapter 7).

[15] Thanks to Rick Lewis for contributing this example.

Appendix A

Contrasting Predictions

The same grammar can lead to different predictions via entropy reduction and surprisal, respectively. This section develops an example where that happens.[15]

probability		rewriting rule		comment
0.98	S	➡	A X	X is radically more probable than Y
0.01	S	➡	B Z	Z is a fairly uncertain category
0.01	S	➡	C Y	choice between X and Y is cued by first symbol, a vs c
1.0	A	➡	a	
1.0	B	➡	b	
1.0	C	➡	c	
1.0	X	➡	f	neither X nor Y is entropic
1.0	Y	➡	g	
0.25	Z	➡	c	
0.25	Z	➡	d	
0.25	Z	➡	e	
0.25	Z	➡	f	

Consider the surprisal value of strings starting with c and b; these are shaded in gray in the first column of the tables below. As specified in Equation 1, this is just the base-2 logarithm of the transition probability, i.e. from 1.0 to 0.01. It is the same across both strings, 6.64 bits. Intuitively, both c and b eliminate the highly probable S → A X rule in favor of one or another lower-probability rule.

The entropy reductions (see Equation 4) associated with these symbols are different: 0.1814 versus none at all. If the first symbol is c, then the rule S → C Y will be required. Since (by construction) there is no uncertainty at all about Y's derivation, this means that all of the entropy associated with S has been reduced.

If instead the first symbol is b, then the rule S → B Z will be required. There is now an 'obligation' to work out Z's derivation. Since Z's derivation is maximally uncertain, entropy goes up: no work is done on this particular symbol. On the following symbol, where Z's derivation is revealed, all of this entropy is reduced.

symbol		**c**	**g**
prefix prob	1	0.01	0.01
surprisal	–	6.64	0
entropy	0.1814	0	0
entropy reduction	–	**0.1814**	0

symbol		**b**	**c**
prefix prob	1	0.01	0.0025
surprisal	–	6.64	2
entropy	0.1814	2.0	0
entropy reduction	–	**none**	2

Bibliography

Attneave, F. 1959. *Applications of information theory to psychology: a summary of basic concepts, methods and results*. Holt, Rinehart and Winston.

Bar-Hillel, Y., M. Perles, and E. Shamir. 1964. On formal properties of simple phrase structure grammars. In *Language and information: selected essays on their theory and application*, 116–50. Reading, Massachusetts: Addison-Wesley.

Blachman, N. 1968. The amount of information that *y* gives about *X*. *IEEE Transactions on Information Theory* 14(1). 27–31.

Boston, M. F., J. Hale, R. Kliegl, U. Patil, and S. Vasishth. 2008. Parsing costs as predictors of reading difficulty: an evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research* 2(1). 1–12.

Boston, M. F., J. Hale, S. Vasishth, and R. Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes* 26, 301–49.

Brants, T., and M. Crocker. 2000. Probabilistic parsing and psychological plausibility. In *Proceedings of 18th international conference on Computational Linguistics COLING-2000*.Saarbrücken/Luxembourg/Nancy.

Brennan, J. R., E. P. Stabler, S. E. VanWaganen, W.-M. Luh, J. T. Hale 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language* 157–158, 81–94. doi:10.1016/j.bandl.2016.04.008

Bresnan, J. 1982. Introduction: Grammars as mental representations of language. In *The mental representation of grammatical relations*, ed. by J. Bresnan, xvii–lii). Cambridge, MA: MIT Press.

Calvillo, J., and M. Crocker. 2015. *A rational statistical parser. In Natural language processing and cognitive science*: *Proceedings 2014* (2015th ed.). Berlin: De Gruyter.

Caplan, D., C. Baker, and F. Dehaut. 1985. Syntactic determinants of sentence comprehension in aphasia. *Cognition* 21(2). 117–75.

Chater, N., A. Clark, A. Perfors, and J. A. Goldsmith. 2015. *Empiricism and language learnability*. Oxford University Press.

Chen, Z., and J. T. Hale. under review. Structural and non-structural uncertainties in expectation-based sentence comprehension.

Chen, Z., T. Hunter, J. Yun, and J. Hale. 2014. Modeling sentence processing difficulty with a conditional probability calculator. In *Proceedings of* 36th *annual cognitive science conference*.

Chomsky, N. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2(3). 113–24.

Demberg, V., and F. Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2). 193–210.

——. 2009. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the 29th meeting of the cognitive science society* (*cogsci-09*) (1888–1893).

Den, Y., and M. Inoue. 1997. Disambiguation with verb-predictability: evidence from Japanese garden-path phenomena. In *Proceedings of the* 19th *annual conference of the cognitive science society* (179–184). Lawrence Erlbaum.

Engbert, R., A. Nuthmann, E. M. Richter, and R. Kliegl. 2005. Swift: a dynamical model of saccade generation during reading. *Psychological Review* 112(4). 777–813.

Ford, M., J. Bresnan, and R. M. Kaplan. 1982. A competence-based theory of syntactic closure. In *The mental representation of grammatical relations*, ed. by J. Bresnan, 727–96. Cambridge, MA: MIT Press.

Frank, S. L. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science* 5(3). 475–94.

Frank, S. L., L. J. Otten, G. Galli, and G. Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language* 140(0). 1–11.

Frazier, L. 1985. Syntactic complexity. In *Natural language parsing: psychological, computational, and theoretical perspectives*, ed. by D. R. Dowty, L. Karttunen, and A. M. Zwicky, 129–89. Cambridge University Press.

Frazier, L., and C. Clifton, Jr. 1996. *Construal*. MIT Press.

Garner, W. 1962. *Uncertainty and structure as psychological concepts*. Wiley.

Gibson, E., L. Bergen, and S. T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110(20). 8051–6.

Goldsmith, J.. 2000. On information theory, entropy, and phonology in the 20th century. *Folia Linguistica* 34(1–2). 85–100.

beim Graben, P., J. D. Saddy, M. Schlesewsky, and J. Kurths. 2000. Symbolic dynamics of event-related brain potentials. *Physical Review E* 62(4). 5518–41.

beim Graben, P., S. Gerth, and S. Vasishth. 2008. Towards dynamical system models of language-related brain potentials. *Cognitive Neurodynamics* 2(3). 229–55.

Grodner, D., and E. Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science* 29, 261–90.

Hale, J. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics*.

—— 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research* 32(2). 101–23.

—— 2006. Uncertainty about the rest of the sentence. *Cognitive Science* 30(4). 609–42.
—— 2011. What a rational parser would do. *Cognitive Science* 35(3). 399–443.
—— 2014. Automaton theories of human sentence comprehension. CSLI Publications.
Hale, J., D. Lutz, W.-M. Luh, and J. Brennan. 2015. Modeling fMRI time courses with linguistic structure at various grain sizes. In Proceedings of the 6th workshop on cognitive modeling and computational linguistics (89–97). Denver, Colorado: Association for Computational Linguistics.
Hawkins, J. A. 2010. Processing efficiency and complexity in typological patterns. In *The oxford handbook of linguistic typology*, ed. by J. J. Song, 206–26. Oxford University Press.
Henderson, J. M., W. Choi, M. W. Lowder, and F. Ferreira 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage* 132, 293–300. doi:10.1016/j.neuroimage.2016.02.050
Hick, W. 1952. On the rate of gain of information. *Quarterly Journal of Experimental Psychology* 4(1). 11–26.
Hockett, C. F.. 1953. Review of C. Shannon and W. Weaver, The Mathematical Theory of Communication. *Language* 29(1). 69–93.
Hume, E., and F. Mailhot 2013. The role of entropy and surprisal in phonologization and language change. In *Origins of sound change: approaches to phonologization*, ed. by A. C. L. Yu, 29–47. Oxford University Press.
Jaeger, T. F. 2010. Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62.
Jelinek, F., and J. D. Lafferty. 1991. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics* 17(3).
Jurafsky, D., and J. H. Martin. 2008. *Speech and language processing*. Prentice-Hall.
Kaplan, R. M. 1972. Augmented transition networks as psychological models of sentence comprehension. *Artificial Intelligence* 3, 77–100.
—— 1995 Three seductions of computational psycholinguistics. In Formal issues in lexical-functional grammar (12), ed. by M. Dalrymple, R. M. Kaplan, J. T. Maxwell, and A. Zaenen. Stanford University CSLI.
Keenan, E. L., and S. Hawkins. 1987. The psychological validity of the accessibility hierarchy. In *Universal grammar: 15 essays*, ed. by E. L. Keenan, 60–85. London: Croom Helm.
Keller, F. 2004. The entropy rate principle as a predictor of processing effort: an evaluation against eye-tracking data. In *Proceedings of EMNLP 2004*, ed. by D. Lin, D. Wu, 317–24. Barcelona, Spain: Association for Computational Linguistics.
King, J., and M. A. Just. 1991. Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language* 30, 580–602.
Kornai, A. 2007. *Mathematical linguistics*. Springer Verlag.
Lehman, J. F., R. L. Lewis, and A. Newell. 1991. *Natural language comprehension in Soar: spring 1991* (2052). CMU.
Levelt, W. J. 1974. *Formal grammars in linguistics and psycholinguistics* (I). Mouton. (Reprinted 2008)
Levy, R. 2008a. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–77.
—— 2008b. A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 13th conference on empirical methods in natural language processing* (234–243). Waikiki, Honolulu.
Lewis, R. L. 1993. *An architecturally-based theory of human sentence comprehension*. Pittsburgh, PA: Carnegie Mellon University.
Lewis, R. L., and S. Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29(3). 375–419.
Lewis, R. L., M. Shvartsman, and S. Singh. 2013. The adaptive nature of eye movements in linguistic tasks: how payoff and architecture shape speed-accuracy trade-offs. *Topics in Cognitive Science* 5(3). 581–610.
Linzen, T., and T. F. Jaeger. 2015. Uncertainty and expectation in sentence processing: evidence from subcategorization distributions. *Cognitive Science*. Still in Online Early View (Online Version of Record published before inclusion in an issue).
Luce, R. D. 2003. Whatever happened to information theory in Psychology? *Review of General Psychology* 7(2). 183–8.
MacDonald, M. C. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes* 157–201.
Malouf, R. forthcoming. *Information theory*. Oxford Bibliographies in Linguistics.
Malsburg, T. von der, R. Kliegl, and S. Vasishth. 2015. Determinants of scanpath regularity in reading. *Cognitive Science* 39(7). 1675–703.
Marr, D. 1982. *Vision: a computational investigation into the human representation and processing of visual information*. W.H. Freeman and Company.
Milin, P., V. Kuperman, A. Kostić, and R. H. Baayen. 2009. Words and paradigms bit by bit: an information-theoretic approach to the processing of inflection and derivation. In *Analogy in grammar form and acquisition*, ed. by J. P. Blevins, J. Blevins, 214–52. Oxford University Press.
Morrill, G. 2000. Incremental processing and acceptability. *Computational Linguistics* 26(3). 319–38.
Nederhof, M.-J., and G. Satta. 2008. Computing partition functions of PCFGs. *Research on Language and Computation* 6(2). 139–62.
Partee, B. H., ter Meulen, A. and R. E. Wall. 1993. *Mathematical methods in linguistics*. Kluwer.

Pierce, J. R. 1980. *An introduction to information theory: symbols, signals & noise*. Dover.

Rauzy, S., and P. Blache. 2012. Robustness and processing difficulty models. a pilot study for eye-tracking data on the French Treebank. In *Proceedings of the workshop on eye-tracking and Natural Language Processing at the* 24th *international conference on computational linguistics* (*COLING*).Mumbai, India.

Roark, B., A. Bachrach, C. Cardenas, and C. Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the conference on empirical methods in natural language processing* (*EMNLP*) (324–333).

Rosenbloom, P., and A. Newell. 1987. Learning by Chunking: a production system model of practice. In *Production system models of learning and development*, ed. by D. Klahr, P. Langley, and R. Neches. MIT Press.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423, 623–656.

Stabler, E. 1994. The finite connectivity of linguistic structure. In *Perspectives on sentence processing*, ed. by C. Clifton, L. Frazier, and K. Rayner, 303–36. Lawrence Erlbaum.

Staub, A. 2010. Eye movements and processing difficulty in object relative clauses. *Cognition* 116(1). 71–86.

—— 2015. The effect of lexical predictability on eye movements in reading: critical review and theoretical interpretation. *Language and Linguistics Compass* 9, 311–27.

Taylor, W. 1953. Cloze procedure: a new tool for measuring readability. *Journalism Quarterly* 30, 415–33.

Thibadeau, R., M. A. Just, and P. Carpenter. 1982. A model of the time course and content of reading. *Cognitive Science* 6, 157–203.

Tribus, M. 1961. *Thermostatics and thermodynamics*. D. van Nostrand Company.

Willems, R. M., S. L. Frank, A. D. Nijhof, P. Hagoort, and A. Bosch, van den. 2015. Prediction during natural language comprehension. *Cerebral Cortex*. (Advance Access published online: April 22, 2015)

Wu, S., A. Bachrach, C. Cardenas, and W. Schuler. 2010. Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (1189–1198). Uppsala, Sweden: Association for Computational Linguistics.

Yngve, V. H. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104(5). 444–66.

Yun, J., Z. Chen, T. Hunter, J. Whitman, and J. Hale. 2015. Uncertainty in processing relative clauses across East Asian languages. *Journal of East Asian Linguistics* 24(2). 113–48.