Journal of East Asian Linguistics manuscript No.
(will be inserted by the editor)

Uncertainty in processing relative clauses

Jiwon Yun · Zhong Chen · Tim Hunter · John Whitman · John Hale

Received: date / Accepted: date

Abstract The processing difficulty profile for relative clauses in Chinese, Japanese and Korean represents a challenge for theories of human parsing. We address this challenge using a grammar-based complexity metric, one that reflects a minimalist analysis of relative clauses for all three languages as well as structure-dependent corpus distributions. Together, these define a comprehender's degree of uncertainty at each point in a sentence. We use this idea to quantify the intuition that people do comprehension work as they incrementally resolve ambiguity, word by word. We find that downward changes to this quantitative measure of uncertainty derive observed processing contrasts between Subject- and Object-extracted relative clauses. This demonstrates that the complexity metric, in conjunction with a minimalist grammar and corpus-based weights, accounts for the widely-observed Subject Advantage.

Keywords relative clause · information theory · minimalism · Chinese · Japanese · Korean

Jiwon Yun
Department of Linguistics
Stony Brook University
E-mail: jiwon.yun@stonybrook.edu

Zhong Chen
Department of Modern Languages and Cultures
Rochester Institute of Technology
E-mail: z.chen@rit.edu

Tim Hunter
Institute of Linguistics
University of Minnesota
E-mail: timh@umn.edu

John Whitman
Department of Linguistics
Cornell University
E-mail: jbw2@cornell.edu

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

Electronic Supplementary Material for this paper is available at URL.

1 Introduction

Relative clauses present linguists with a variety of puzzles. Two in particular are fundamental in the sense that any solution to them would carry implications for syntax, typology and psycholinguistics. The first is their structure. Can an analysis be given that is simultaneously general, consistent and descriptively accurate across multiple languages? The second is their processing. Why are some relative clauses easier to understand than others? Can asymmetries in processing be related to their syntactic structure, and if so how? This paper proposes a solution to these puzzles. The basic idea is that subject and object relatives have different amounts of sentence-medial ambiguity. We show that a quantitative version of this idea can account for the observed processing asymmetry between these two sentence types.

To show this, we use Minimalist Grammars (Stabler, 1997) to explicitly define the set of alternative syntactic analyses that are consistent with the sentence-initial substring of words perceived so far. Probability distributions over this *remainder* set characterize the expectations that a comprehender would have, partway through the sentence. To estimate them, we weight the grammars' structural rules in accordance with corpora like the Penn Chinese Treebank 7 (Xue, Xia, Chiou, and Palmer, 2005), the Kyoto Corpus 4 (Kurohashi and Nagao, 2003) and the Penn Korean Treebank 2 (Han, Han, Ko, and Palmer, 2002). The uncertainty or *entropy* of this remainder set quantifies the notion of sentence-medial ambiguity. If this entropy goes down in the transition from one word to the next, then sentence-processing work must have occurred (Wilson and Carroll, 1954; Hale, 2006).

These assumptions suffice to derive a universal Subject Advantage in processing. Keenan and Hawkins (1974/1987) found early support for such a universal, and since then a growing body of experimental work has supported it. Proceeding from these findings, this paper fleshes out a role for minimalist syntax in a quantitative theory of derivational uncertainty. In particular, the proposed theory

uses syntactic alternatives that are logically entailed by the grammar to derive numerical predictions of processing difficulty at specific words.

These predictions, across Chinese, Japanese & Korean (henceforth: CJK) are laid out in section 6. But before that, section 2 first reviews the relevant psycholinguistics literature. Section 3 goes on to discuss the syntactic analysis of relative clauses involving subject and object gaps. For such constructions, there is a fairly strong consensus in the generative literature that some kind of extraction is implicated in the syntactic derivation (see Huang, Li, and Li (2009) for Chinese, Whitman (2012) for Japanese and Han and Kim (2004) for Korean). Section 4 introduces relative frequency estimation, a simple way of using corpora to put weights on grammar rules. This weighting makes it possible to quantify sentence-medial ambiguity as the entropy of the remainder set. Section 5 reviews the Entropy Reduction hypothesis, which links fluctuating entropy levels and processing difficulty. Finally, section 7 concludes with a reflection on human sentence processing as information processing. Our grammars and corpus data are included as Electronic Supplementary Material, and our software is freely available.

2 Asymmetries in Processing Relative Clauses

2.1 Relative Clauses and the Subject Advantage

In a relative clause (RC), a noun phrase is said to have been "relativized" from one of a number of different "underlying" positions, for example subject position or object position. The RC construction as a whole thus exhibits a Filler–Gap relationship. A large literature documents the finding that subject relatives (SRCs) are consistently easier to process than object relatives (ORCs) across languages, a processing asymmetry often referred to as the Subject Advantage. This Subject Advantage has been observed for English using a variety of different measures, including: reading times (King and Just, 1991), eye-tracking (Traxler, Morris, and

Seely, 2002), ERP (King and Kutas, 1995), fMRI (Just, Carpenter, Keller, Eddy, and Thulborn, 1996) and PET (Stromswold, Caplan, Alpert, and Rauch, 1996). European languages other than English also attest the Subject Advantage, for instance Dutch (Frazier, 1987; Mak, Vonk, and Schriefers, 2002), French (Frauenfelder, Segui, and Mehler, 1980) and German (Schriefers, Friederici, and Kühn, 1995; Mecklinger, Schriefers, Steinhauer, and Friederici, 1995).

In all these languages, the RC appears after the noun it modifies. That is to say: English, French, German and Dutch all have postnominal RCs. By contrast, CJK RCs come before the noun they modify. This prenominal positioning of the RC with respect to the *head noun* is illustrated below[1] in examples 1–3.

(1) Chinese

a. SRC

[e_i shushi furen de] (jingli$_i$)
knows tycoon DE manager/someone
'the manager/someone who knows the tycoon'

b. ORC

[furen shushi e_i de] (jingli$_i$)
tycoon knows DE manager/someone
'the manager/someone who the tycoon knows'

(2) Japanese

a. SRC

[e_i daigisi o hinansita] kisya$_i$
senator Acc criticize reporter
'the reporter who criticized the senator'

b. ORC

[daigisi ga e_i hinansita] kisya$_i$
senator Nom criticize reporter

[1] Parentheses in example 1 indicate the optionality of the head noun in Chinese. The other abbreviations in the glosses are:
Nom nominative case marker
Acc accusative case marker
Decl declarative verb ending
Adn adnominal verb ending
See also Table 4 on page 22.

'the reporter who the senator criticized'

(3) Korean

a. SRC

[e_i uywon ul pinanhan] kica$_i$
senator Acc criticize.Adn reporter
'the reporter who criticized the senator'

b. ORC

[uywon i e_i pinanhan] kica$_i$
senator Nom criticize.Adn reporter
'the reporter who the senator criticized'

The distinction between prenominal and postnominal positioning makes CJK RCs a uniquely valuable domain for testing universalist claims about human sentence processing. Below, we summarize some key empirical findings. Section 2.3 then goes on to assess the available theories.

2.2 The Subject Advantage in CJK

A variety of experiments have measured the Subject Advantage at specific points in sentences containing prenominal relative clauses. For instance, using self-paced reading, Ishizuka, Nakatani, and Gibson (2003) and Miyamoto and Nakamura (2003, 2013) found it at the head noun in Japanese RCs. Ishizuka (2005) reports it as well at the RC-initial case-marked noun phrase. Korean is much the same: Kwon, Polinsky, and Kluender (2006), for instance, reports a Subject Advantage at the head noun. These contrasts have been replicated across several different methodologies including ERP in Japanese (Ueno and Garnsey, 2008), and eye-tracking in Korean (Kwon, Lee, Gordon, Kluender, and Polinsky, 2010).

The processing of Chinese RCs, on the other hand, has been harder to pin down. Early work by F. Hsiao and Gibson (2003) reported the inverse result, an Object Advantage, in contrast to the Subject Advantage found in Lin and Bever

(2006). Further analysis indicates that this outcome may have been due to uncontrolled factors, such as local ambiguities (C. Lin and Bever, 2011; Qiao, Shen, and Forster, 2012; Vasishth, Z. Chen, Li, and Guo, 2013; Y. Hsiao, Li, and MacDonald, 2014)[2] and syntactic priming from the context with different thematic orders (Lin, 2014). With stimuli that control local ambiguities such as the availability of argument omission, Jäger et al. (submitted) observe a robust Subject Advantage in Chinese RCs. A wider array of references is provided below in Table 1. Overall, the weight of the evidence seems to suggest that Chinese is not exceptional after all but rather confirms the universalist view: the Subject Advantage manifests itself in both prenominal and postnominal RCs.

Subject Advantage	**Object Advantage**
C. Lin and Bever (2006, 2007, 2011)	F. Hsiao and Gibson (2003)
C. Lin (2008; submitted)	B. Chen, Ning, Bi, and Dunlap (2008)
F. Wu (2009)	Y. Lin and Garnsey (2011)
F. Wu, Kaiser, and Andersen (2012)	Packard, Ye, and Zhou (2011)
Vasishth et al. (2013)	Qiao et al. (2012)
Jäger et al. (submitted)	Gibson and H. Wu (2013)
F. Wu and Kaiser (submitted)	

Table 1 Psycholinguistic experiments of Chinese relative clauses

2.3 Theories of Relative Clause Processing

A variety of general principles have been advanced to account for the Subject Advantage. Table 2 catalogs some of the leading ideas. Among these, recent work has been especially concerned with Memory Burden and Structural frequency. The memory burden idea, while appealing for European languages, does not work for CJK RCs that are prenominal. To see this, consider the distance between the head noun and its coindexed empty category e_i in any of the examples 1–3. In

[2] These sorts of ambiguities represent an exciting research area for CJK psycholinguistics. For a review with special emphasis on Japanese, see Hirose (2009).

Broad Categories		General Proposals
Word Order	Bever (1970); MacDonald and Christiansen (2002)	The sequence of words in SRCs is closer to the canonical word order than that in ORCs.
Parallel Function	Sheldon (1974)	SRCs are easier to process than ORCs because their head nouns play the same role in both the main clause and the subordinate clauses.
Perspective Maintenance	MacWhinney (1977, 1982)	SRC structures maintain the human perspective and should be easier to process than those that shift it, e.g. ORCs.
Accessibility Hierarchy	Keenan and Comrie (1977)	Universal markedness hierarchy of grammatical relations ranks the relativization from subject higher.
Memory Burden	Linear Distance: Wanner and Maratsos (1978); Gibson (2000); Lewis and Vasishth (2005)	ORCs are harder because they impose a greater memory burden.
	Structural Distance: O'Grady (1997); Hawkins (2004)	
Structural Frequency	Tuning Hypothesis: Mitchell, Cuetos, Corley, and Brysbaert (1995); Jurafsky (1996)	SRCs occur more frequently than ORCs and therefore are more expected and easier to process.
	Surprisal: Hale (2001); Levy (2008)	ORCs are more difficult because they require a low-probability rule.
	Entropy Reduction: Hale (2006)	ORCs are harder because they force the comprehender through more confusing intermediate states.

Table 2 Processing principles proposed for relative clauses

all cases, the predicted memory burden in the SRC would be greater than in the ORC, contrary to the observed empirical pattern.

The structural frequency idea is also appealing, since by and large the attestation rate of SRCs exceeds ORCs. However, in its most well-known incarnation as "surprisal" this idea also fails to derive the observed data. Levy (2008) acknowledges the situation when he writes

> 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 [...]
>
> page 1166

At least in Levy's formulation, surprisal does not work for English RCs. Seeking a more adequate complexity metric, Hale (2003, 2006) advances an alternative called Entropy Reduction. Like surprisal, Entropy Reduction is information-theoretical. But unlike surprisal it correctly derives the observed Subject Advantage in English. Section 2.4 reviews the original account given on pages 116–118 of Hale (2003) in light of subsequent work.

2.4 Entropy Reduction as a complexity metric

The basic idea of Entropy Reduction is that comprehenders struggle against the tide of ambiguity that they face in the course of incremental processing. Words, as they come in, are either helpful or unhelpful in narrowing down the interpretation that the speaker (or writer) intends. As in Hale (2006), we consider intermediate stages that correspond to sentence-initial substrings. Example 4 illustrates these stages with an example used in the account of English RC processing difficulty from that paper.

(4) initial substrings of "the sailor who s ship Jim take -ed have -ed one leg"

a. (empty string)

b. the

c. the sailor

d. the sailor who

e. the sailor who s

⋮

The symbols in 4 mostly correspond to whole words, although some morphemes such as the genitive *s* get their own symbol. From this perspective, the latest symbol added to the initial substring is taken to be 'informative' about the overall structure of the unfolding sentence. This contribution is quantified by changes in the conditional entropy of the derivation given the initial string.

The basic idea has a long history. To the best of our knowledge, it was introduced in section 5.3 of Wilson and Carroll (1954). At this time, many cognitive scientists were interested in applying information theory to human communication (see for example chapter 5 of Cherry (1961), section 6.1 of Levelt (1974, volume II), or chapter 3 of Smith (1973)). Wilson and Carroll applied Entropy Reduction to an artificial language of their own creation, by way of introducing the idea and demonstrating its potential utility for morphosyntactic analysis. In doing so, they also acknowledged a major restriction: their formulation relied on a Markov model of language, analogous to beads-on-a-string.

Hale (2003) revived the Entropy Reduction idea by lifting this restriction and applied it to the analysis of processing asymmetries in English, including the Subject Advantage. This paper used context-free phrase structure grammars as models of language structure. The account of the Subject Advantage essentially turned on the possibility of recursive modification which exists in the SRC but which is eliminated by the embedded verb in the ORC. Later papers such as Hale (2006) upgraded the language model yet further to expressive formalisms like Stabler's Minimalist Grammars (1997) where a movement analysis of relativization can be stated directly. This change was accompanied by new algorithms for computing the metric, but the basic equation between human sentence processing work and the reduction of derivational uncertainty remained the same.

Section 9.1 of Hale (2006) notes that Entropy Reduction can derive the repetition accuracy cline that is observed along Keenan and Comrie's (1977) well-known Accessibility Hierarchy of relativizable grammatical relations (AH). This same pattern is also part of the empirical support for the Minimize Domains (MiD) principle (Hawkins, 2004, §7.2). Entropy Reduction and MiD both derive the AH but from different starting points. Whereas MiD (and its predecessor, Early Immediate Constituents) considers the number of syntactic nodes involved in the ultimately-correct analysis of e.g. an English RC, Entropy Reduction takes into account changing distributions on intermediate parser states. It quantifies the idea

of sentence-medial ambiguity with a numerical uncertainty level over intermediate parser states.

There is no necessary connection between Entropy Reduction as a complexity metric and Minimalist Grammars as a formalism. In fact, Frank (2013) recently applied Entropy Reduction to the analysis of British readers' eye fixation times using Simple Recurrent Nets as substitute for a grammar. In this study, Entropy Reduction emerged as a significant predictor of fixation duration. Interestingly, as he computed Entropy Reductions with greater and greater fidelity, Frank found that their fit to observed fixation times got better and better. This sort of result strengthens our confidence in the metric itself.

What remains to be shown is that Entropy Reduction derives the correct predictions in the key domain of prenominal RCs. This is important because, as things stand, there exists no formalized account of incremental processing difficulty that accords with data on both prenominal and postnominal RC processing. In order to make these predictions, we first have to fix upon a syntactic analysis. The syntactic analysis, discussed in section 3, motivates a particular series of corpus studies (section 4) aimed at weighting the rules of a formalized grammar fragment.

3 Structure of Relative Clauses

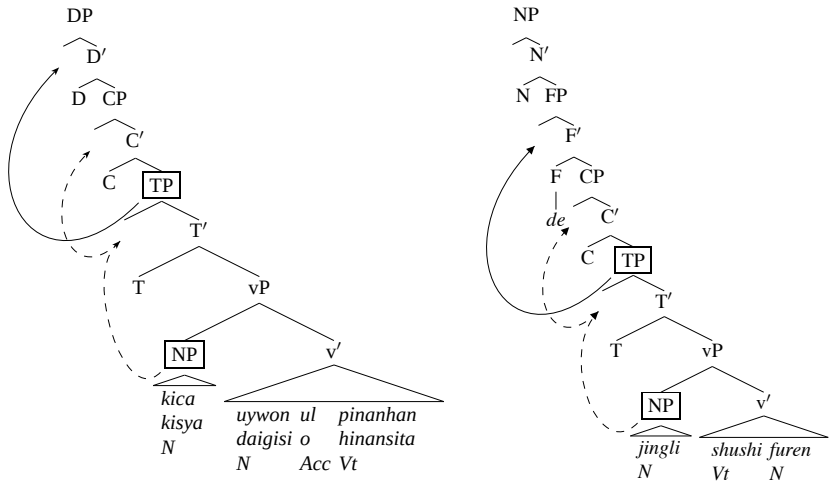
As noted in the Introduction, there is a general consensus in the generative literature that movement is involved in the derivation of at least some CJK RCs. This evidence comes from the existence of island effects as well as subtler effects such as scope reconstruction and idiom chunk effects (Huang et al., 2009; Whitman, 2012; Han and Kim, 2004). Notice, however, that the consensus claim is not exhaustive. Other RCs seem to present island violations. A widespread view holds that these are cases of resumptive *pro*, where movement is blocked. But we restrict ourselves here to Subject and Object gaps in non-island environments. In these cases a movement analysis is appropriate.

3.1 The promotion analysis

There are two main movement analyses of RCs in the generative literature. An analysis moving a null operator from the position of the gap into the clausal projection of the RC was popular through the 1990s (Ning (1993) for Chinese, Ishii (1991) for Japanese, Kaplan and Whitman (1995) for Japanese and Korean). Kayne (1994) revives a different movement analysis for RCs, dating back to Brame (1967), Schachter (1973) and Vergnaud (1974). Under this "promotion" analysis, the RC head is moved directly from the position of the gap into RC head position. Kayne points out that this analysis is particularly attractive for prenominal RCs because it explains the well-known typological fact that prenominal relatives never include overt relative pronouns at the beginning of the RC. According to Kayne's analysis, this is because the RC head is first moved from TP to the specifier of CP. TP is then fronted around the head. The derivation thereby accounts for the head/RC order, the absence of relative pronouns at the beginning of the RC, and the absence of evidence that prenominal RCs are CPs, rather than simple TPs. A further, theory-internal advantage of the promotion analysis is that it avoids a violation of the Extension Condition (Chomsky, 1993) inherent in at least some operator movement analyses. Under such analyses, the RC is adjoined to the head of the nominal projection (or in some analyses, generated in Spec, DP). Relative operator movement then takes place within the CP, but this movement is not at the edge (the structurally highest position) inside the nominal projection. On the promotion analysis, by contrast, the NP generated in the position of the gap moves consistently to the edge of the projection at each step of the derivation.

Promotion analyses have been adopted by X. Wu (2000) for Chinese, and Hoshi (1995) for Japanese. Huang et al. (2009) adopt a mixed analysis of Chinese RCs, involving both promotion and operator adjunction. As our purpose in this paper is to generalize across comparable derivations, in the formal grammar fragments, we implemented promotion derivations, as representative of the currently most widely-adopted analysis of RCs in minimalist theory. Figure 1 sketches these

derivations with Korean and Japanese presented at the same time in 1(a) and Chinese separately in 1(b).



(a) Korean and Japanese (b) Chinese

Fig. 1 Syntactic analyses of CJK relative clauses

3.2 Minimalist Grammars

The promotion analysis discussed above in section 3.1 is a compelling idea about the general structure of RCs across languages. But in order to extract specific quantitative predictions, one must express this general idea in some concrete way. In this project, we did this using the Minimalist Grammars (MGs) of Stabler (1997). This system formalizes certain themes of Chomsky's Minimalist Program (1995). For instance, MG lexical entries have sequences of features that motivate derivational operations. One kind of feature motivates the MERGE operation; this aspect is reminiscent of categorial grammar (Berwick and Epstein, 1995). Another kind of feature motivates the MOVE operation; this operation reorganizes the derived tree and has no direct parallel in categorial grammar. In our grammars, for instance, there is a `+wh` feature that motivates movement of a corresponding `-wh` phrase.

The finitude of the available feature types, in any given grammar, is crucial for reining in the expressive power of these grammars.

Minimalist Grammars can be viewed as defining two operations (MERGE, MOVE) in a kind of bottom-up way. But this perspective is not exclusive. An important mathematical result shows that the same formalism can also be viewed as a rewriting system that works top-down (Michaelis, 2001; Harkema, 2001). Neither direction is privileged if one adopts a view of MG derivations as trees. One can get a sense of these *derivation trees* by imagining nodes labeled with the operation name (MERGE or MOVE) and leaves labeled with feature bundles.[3]

The derivation tree viewpoint underwrites the interpretation of MGs as stochastic branching processes, and therefore also their interpretation as probabilistic grammars.[4] The core idea is that a derivation may continue, from the top down, by application of any one of a number of alternative operations. These operations are "backwards" from the usual bottom-up perspective but this presents no difficulty. If appropriately normalized, the weights on these alternative branches become probabilities and the grammars themselves can take on a more cognitive interpretation: they define the class of structures a comprehender might be expecting.[5] In order to set up this cognitive interpretation (section 5), numerical weights have to be determined. Section 4 explains how we used corpus distributions to obtain these numbers.

[3] See Figure 12 of Hale (2006). The distinction between derivation tree and derived tree was introduced into cognitive science by Aravind Joshi.

[4] Chapter 2 of Harris (1963) treats, in detail, the sort of branching process used in these linguistic models. Smith (1973, pages 66-68) and Hale (2006, §3.1) both review the work of Grenander (1967) who was the first to see the connection to branching processes.

[5] The modeling in this paper relies exclusively on syntactic information e.g. grammatical category, hierarchical phrase structure and WH-movement. This leaves out nonsyntactic factors, such as animacy and information structure, which also play a crucial role in human sentence comprehension. However, nothing in the overall approach prevents inclusion of additional features in the formal grammar fragment e.g. diacritics such as ANIMATE or TOPIC. In continuing modeling work, Z. Chen (2014) estimates weighted grammars with formalist as well as "functionalist" feature names. The results accurately reflect the Subject Advantage and the animacy effect in English, Italian and Chinese RCs.

4 From corpora to weighted grammars

The step from corpora to weighted grammars follows a simple logic: branches of the derivation tree represent choices about which structure to generate. In a performance model, these choices might reasonably be based on experience. We can approximate this experience by estimating various statistics from Treebanks or other samples. This section illustrates one such procedure, starting from a simplified example based on context-free grammars. It turns out, however, this is not really a simplification at all, because the weighting of MG derivations proceeds in exactly the same manner.

4.1 Relative frequency estimation for tree branches

Because the fragments of Chinese, Japanese and Korean are expressed as MGs, their derivations may be viewed as having been generated by a context-free phrase structure grammar (CFG). The estimation problem therefore reduces to the problem of weighting a CFG. But this problem is easy to solve; the method can be demonstrated with a small example such as Figure 2. The rules in Figure 2 present

S	→	NP VP
NP	→	Det N
NP	→	N
VP	→	V
VP	→	V NP

Fig. 2 Unweighted grammar offers alternative ways to rewrite both NP and VP

us with choices: is an NP going to have a determiner (Det) or will it be a bare noun (N)? Similarly, will a verb phrase be transitive or intransitive? The idea of relative frequency estimation is to set the weights on these choices according to the ratio of those two structure types in a sample. This task is far easier if one has access to a Treebank — a corpus whose sentences are annotated with phrase structures.

Given a Treebank, one can easily weight this grammar by counting. First, count how many times an NP node appears with two daughters, one labeled Det and one labeled N; say this number is 100. Second, count how many times an NP node appears with a single N daughter; say this number is 50. Similarly, suppose that a VP node appears with a single V daughter 120 times, and that a VP node appears with a V daughter and an NP daughter 90 times. These counts can be summarized as shown in (5).

(5)

NP with two daughters, labeled Det and N	100
NP with a single daughter, labeled N	50
VP with a single daughter, labeled V	120
VP with two daughters, labeled V and NP	90

Then we would assign the 'NP $\rightarrow$ Det N' rule a weight of $\frac{100}{150}$ (or $\frac{2}{3}$), and the 'NP $\rightarrow$ N' rule a weight of $\frac{50}{150}$ (or $\frac{1}{3}$). Similarly, the two VP rules would be assigned weights of $\frac{120}{210}$ and $\frac{90}{210}$ respectively. Note that when a nonterminal, such as S in this example, has only one possible expansion, then this rule must have a weight of one and so there is no need to consult a corpus.

The situation is analogous for the more complex MG rewriting system: counts of how often certain grammatical patterns appear in corpora still suffice to determine the relevant weights. To illustrate, the relevant corpus frequencies for our Korean MG are given in (6). For simplicity, we omit from (6) the details of the rewriting system's parent-daughter combinations that are indicated explicitly in Figure 2 and (5), and instead describe the relevant grammatical patterns in formalism-neutral terms (e.g. "intransitive verb", "complement clause"). The key point is that these counts determine the weights for our Korean grammar in precisely the same way that the structural counts in (5) determined the weights for the grammar in Figure 2. As in Figure 2 the counts are grouped into alternatives; each decision in the stochastic derivation process requires choosing from one of these pairs. Setting weights for each of the two alternatives at each such

choice point fully determines a probability distribution over the derivations of our grammars, among which are the RC derivations discussed in section 3.

(6)

intransitive verb	1417
transitive verb	1038
pro in subject position	594
non-*pro* noun phrase in subject position	961
pro in object position	23
non-*pro* noun phrase in object position	1015
subject relative clause	1030
object relative clause	130
relative clause	1160
complement clause	902
noun phrase with complement clause or relative clause	2062
noun phrase consisting of only a simple noun	1976

Although MGs differ from CFGs (e.g. by involving movement operations) counts of how often certain grammatical patterns appear in corpora still suffice to determine the weights of the relevant grammatical decisions. For example, note that one of the decisions that appears in our Korean grammar is the choice between SRCs and ORCs. While this can be thought of as choosing a gap position, in an MG this decision amounts to the choice between two different node labels on the same derivation tree node. In this more complex case, we again compute weights simply by counting how many SRCs and how many ORCs appear in a corpus and normalizing appropriately.

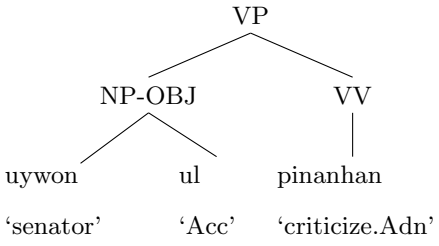
4.2 Counting constructions in corpora

To see how this counting is accomplished, this section offers two examples. These examples are from Korean, although the procedure for the Japanese and Chinese grammars is very similar. The chief consideration to keep in mind is the distinction between the formalism in which the Treebank is encoded (typically phrase

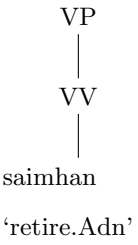
structure) and the formalism being applied in the linguistic performance model (here, MGs).

The first example is transitivity. How can we estimate the rate at which VPs are transitive as opposed to intransitive? Looking in the Penn Korean Treebank (KTB) (Han et al., 2002) we identify particular structural configurations that the annotators used to flag these alternatives.[6] These configurations are schematically illustrated below in 7a and 7b.

(7) a. transitive configuration



b. intransitive configuration



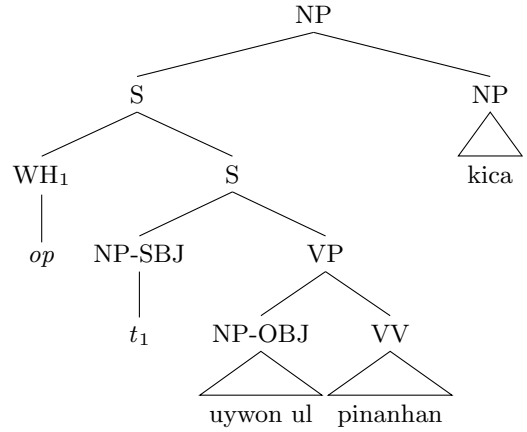
The count of configuration 7a (C_a) compared to configuration 7b (C_b) estimates the transitivity parameter that we would need as a weight in our grammar. In a grammar like the one in Figure 2 we would assign $\frac{C_a}{C_a+C_b}$ to the transitive VP rule ‘VP $\rightarrow$ V NP’ and $\frac{C_b}{C_a+C_b}$ to other, intransitive rule ‘VP $\rightarrow$ V’.

The second example is relativization from Subject as opposed to Object. Here the procedure is exactly the same. The KTB represents SRCs as an S node which (i) adjoins to an NP, and (ii) has as its first daughter a WH-operator that is coindexed with a trace in an NP-SBJ position; similarly for ORCs and an NP-OBJ position. Specific cases of these two patterns are shown below in 8a and 8b. Of course, the criterial features are the tree configurations and the coindexation

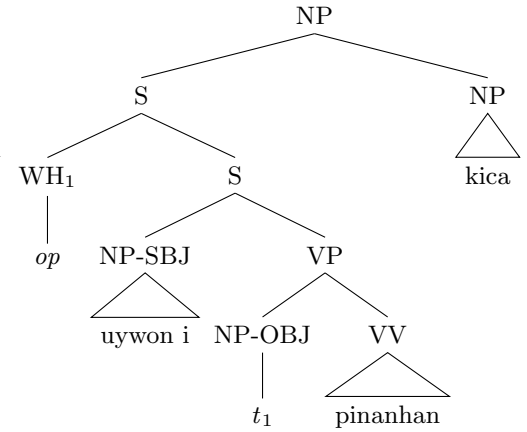
[6] We employ the pattern matching tool Tregex (Levy and Andrew, 2006) for our corpus search.

– not the specific words. These structures are the ones that qualify as SRCs or ORCs in the sense required for the totals shown in (6).

(8) a. SRC configuration



b. ORC configuration



Appendix A includes more details e.g. about weighting the Japanese and Chinese grammars. The logic of the approach is identical, but we have restricted attention to Korean here because certain technical details make it the simplest of the three cases.

5 Weighted grammars and information gain

The Entropy Reduction hypothesis, referred to above in section 2.4, requires some sort of weighting to quantify degrees of expectation. Section 4 introduced the weight-setting methodology. But equally required for a solid understanding of

Entropy Reduction is the idea of conditioning grammars on initial substrings of sentences. This aspect, taken up in this section, is symbolized by the vertical dimension in Table 3.

(Generative) Grammar discrete formal system generates a sentence if a derivation exists	Weighted Grammar discrete formal system with weights generates (sentence,weight) pairs sentence-weights accumulate rule-weights
Intersection Grammar generates a sentence if a derivation exists subject to a condition on its yield	Weighted Intersection Grammar generates a (sentence,weight) pair if a derivation exists and the yield-condition is satisfied

Table 3 Generative grammar and two augmentations of it

Entropy Reduction is an *incremental* complexity metric; this means that it makes predictions on a word-by-word basis. These predictions reflect a quantity – entropy – calculated over a set of derivations that remain "in play" at each successive word in a sentence. If no words have been heard, then the set of in-play derivations is, of course, the full set of all possible derivations. But as the comprehender considers successive words of an initial substring, these words begin to impose more of a constraint. With an explicit grammar fragment, we can calculate the *remainder set* of allowable word-sequences that can grammatically follow some given initial substring. We call members of this set "remainders." The remainders are the strings that would be generated by a new grammar representing the intersection of the given initial string with the original, full-sentence grammar (Bar-Hillel et al., 1964). This intersecting, in the style of Bar-Hillel, is indicated in the transition from the upper-left quadrant to the lower-left quadrant of Table 3. Each remainder has at least one derivation associated with it, and these derivations are compactly encoded by what is called an intersection grammar. In this paper, the yield condition is that the first few words must match an explicitly-given list representing the words that have already been heard or read, as in example 4. We call

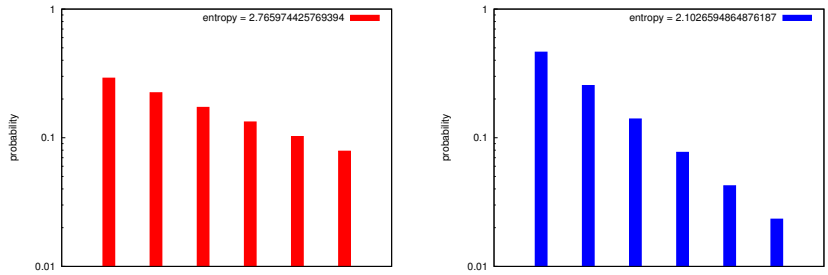
intersection grammars meeting this condition "remainder grammars." Transitioning to the bottom-right quadrant of Table 3, we add weights. This change does not alter the requirement that each derivation in the remainder grammar should remain consistent with the same initial string. But importantly, the weights can now quantify the degree to which an idealized comprehender would expect one remainder or another.

The information gained from a word in a sentence is precisely the drop in derivational uncertainty that is precipitated by that word. This uncertainty is formalized using the definition of entropy (H) given below in 9.

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i) \tag{9}$$

In this application to performance models of language, the $p(x_i)$ in definition 9 are the probabilities of syntactic derivations that are still "in play" at a given initial substring. Using the methods of Grenander (1967) and Nederhof and Satta (2008), it causes no particular difficulty if the number of these derivations is unbounded (see Appendix B).

Weighted grammars thus define a probability distribution on what might be coming up later in the sentence, assuming it turns out to be grammatical. If we graph this probability distribution in a way such that each remainder-derivation x_i has its own bar and the height of the bars correspond to their probability $p(x_i)$ then we can interpret the entropy of this distribution visually as the flatness of the graph. Given two distributions over the same events, the distribution with the flatter graph is the *higher-entropy* distribution. Figure 3 illustrates this sort of comparison with two distributions with geometrically-decreasing probabilities, just like those defined by the weighted grammars for CJK. By visual inspection, one can see how the lower entropy distribution 3(b) concentrates its probability on the left-hand side of the graph more heavily than the higher entropy distribution 3(a) does, the latter being more spread out. The maximum entropy distribution



(a) Entropy = 2.8 bits (b) Entropy = 2.1 bits

Fig. 3 Two distributions, one of which is more peaked.

is of course the uniform distribution which, if graphed, would be a flat horizontal line.

If a particular word ends up tightening the overall constraints on the remainder set, then entropy has been reduced. Ultimately, if the sentence is unambiguous then uncertainty about the intended derivation should be brought down to zero by the end of the utterance. But locally, words that open up more options than they close down can cause entropy to increase. In some interpretations of information theory this is interpreted as "negative information" however, as its name implies, Entropy Reduction only considers transitions on which information is actually gained. Appendix C goes into additional aspects of information theory.

6 Processing Predictions

This section reports the incremental processing difficulty profiles that derive from the syntactic analysis discussed in section 3 via the Entropy Reduction complexity metric reviewed in subsection 2.4. This general methodology is discussed at length in Hale (2006). In all three languages the pattern is the same: a Subject Advantage is derived such that SRCs are correctly identified as easier to understand. The following subsections discuss in detail the positions at which these Subject Advantages are predicted — typically at the beginning of the relative clause region and at its head noun. The discussion relates these predicted incremental processing

Descriptive labels for terminal nodes	
N	noun
V	verb
	-t: transitive -i: intransitive
	-d: declarative -n: adnominal
fact	nouns such as *fact* that take a complement clause
de	relativizer in Chinese
Nom	nominative case marker
Acc	accusative case marker
Dem	demonstratives
Cl	classifiers
Time	temporal phrases

Symbols appearing in derived strings	
t	trace; indicates movement's launching site
pro	unpronounced pronoun; not derived by movement
e	empty category, unspecified
[ ]	brackets indicate embedding

Syntactic Constructions	
SRC	Subject Relative Clause
ORC	Object Relative Clause
NCC	Noun Complement Clause

Linking Hypothesis	
ER	Entropy Reduction

Table 4 Abbreviations

asymmetries to grammatical alternatives defined by a formal grammar fragment. We focus on the role of sentence structure in processing by using grammar fragments that derive part-of-speech symbols, rather than particular lexical items. A complete list of these symbols is provided in Table 4.

6.1 Korean

The word-by-word difficulty predictions for Korean RCs using Entropy Reduction are shown in Figure 4. This Figure illustrates that more processing effort is predicted for ORCs than SRCs in general. Specifically, a Subject Advantage stands out on the second word (i.e. case marker) and the fourth word (i.e. head noun). The rest of this section details the derivation of these processing asymmetries at these specific positions.

Figure 5 tracks how the most probable structures and their probabilities change as the comprehender encounters new words. Each box lists sentence-level strings generated by the remainder grammar at the given initial substring, highlighted in bold, along with the entropy of the probability distribution over those remain-

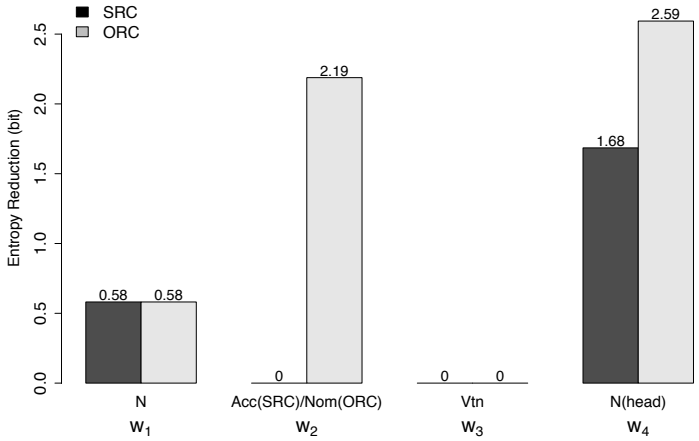


Fig. 4 ER prediction for Korean RCs

ders. These entropy values assist in the interpretation of the entropy reductions (ER) shown on each arrow, which are the actual difficulty predictions graphed in Figure 4. The strings also include non-boldface words that are anticipated but as yet unheard. Each alternative is ranked by the conditional probability of the whole structure given the common initial substring. For instance, when readers have only encountered the noun symbol N as the first word (w_1), the most likely full sentence is a simple intransitive clause, *N Nom Vid* (probability 0.563).

The Subject Advantage at the case marker (w_2) can be explained as follows. More processing difficulty is predicted for ORCs at w_2 because remainder-set entropy is reduced in the transition from w_1 to w_2 when processing the ORC (ER 2.18 bits), whereas there is no reduction in the case of the SRC. Since both SRC and ORC start with the same word at w_1 (i.e. N), the prediction rests on the difference in the entropy value at w_2. In other words, the Subject Advantage at w_2 is due to the lower entropy of the ORC-initial substring *N Nom* (4.210 bits), compared to the SRC-initial substring *N Acc* (11.073 bits). This difference in entropy can be understood by reference to remainder sets at w_2, as shown in Figure 6. This Figure illustrates that while *N Nom* is very likely to be a matrix subject (0.713), the remainder distribution at *N Acc* is less concentrated. Therefore, the ORC-initial substring at w_2 is associated with a lower entropy value than the cor-

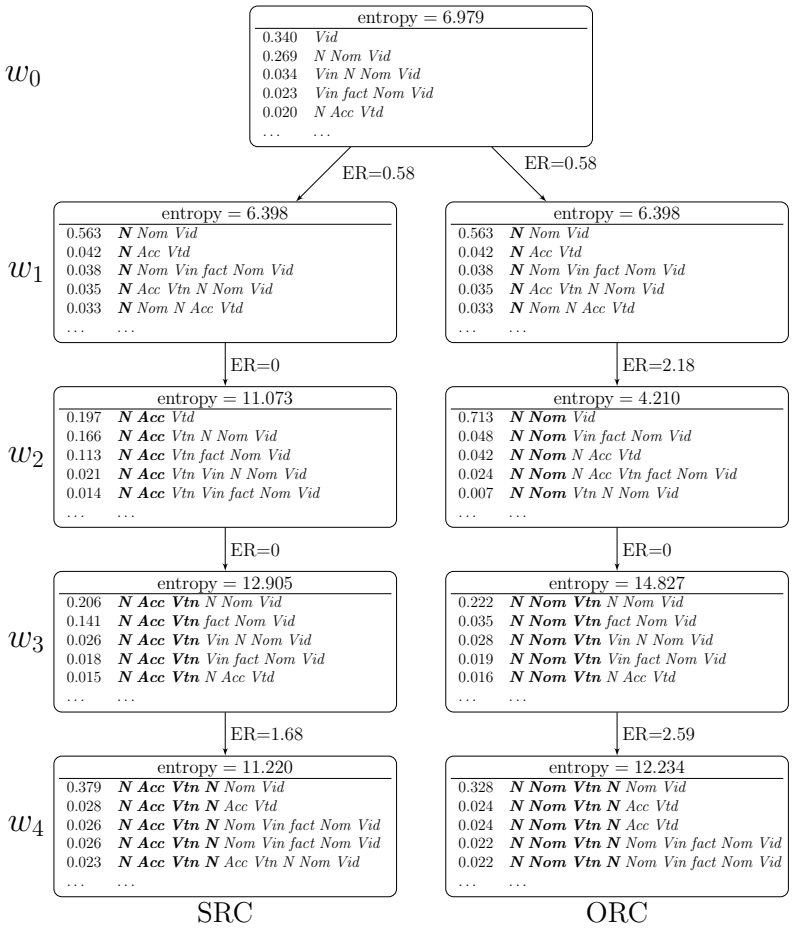


Fig. 5 Likely derivations and their conditional probabilities in Korean RCs

responding SRC-initial substring, which leads to more Entropy Reduction in the ORC's transition from w_1 to w_2.

The Subject Advantage at the head noun (w_4) is also indicated by greater Entropy Reduction for the ORC in the transition from w_3 to w_4. In this transition, the remainder-set entropy is reduced more for the ORC (ER 2.59 bits) than for the SRC (ER 1.68 bits). At w_3, the entropy is higher for the ORC (14.827 bits) than for the SRC (12.905 bits). Once the head noun is revealed at w_4, the entropy is still higher for the ORC but the difference in entropy between the two states becomes relatively small (11.220 bits for SRC and 12.234 bits for ORC). Thus,

Probability	Remainder	Type
0.197	*pro* ***N Acc*** *Vtd*	simplex SOV with Sbj-*pro*
0.166	[*t* ***N Acc*** *Vtn*] *N Nom Vid*	SRC
0.113	[*pro* ***N Acc*** *Vtn*] *fact Nom Vid*	NCC with Sbj-*pro*
0.021	[*t* ***N Acc*** *Vtn*] [*t Vin*] *N Nom Vid*	stacked SRCs
0.014	[*t* ***N Acc*** *Vtn*] [*pro Vin*] *fact Nom N Acc Vtd*	SRC / NCC with Sbj-*pro*
...	...	...
entropy = 11.073		

(a) SRC-initial substring *N Acc*

Probability	Remainder	Type
0.713	***N nom*** *Vid*	simplex SV
0.048	[***N nom*** *Vin*] *fact Nom Vid*	NCC
0.042	***N nom*** *N Acc Vtd*	simplex SOV
0.024	[***N nom*** *N Acc Vtn*] *fact Nom Vid*	NCC
0.007	[***N nom*** *t Vtn*] *N Nom Vid*	ORC
...	...	...
entropy = 4.210		

(b) ORC-initial substring *N Nom*

Fig. 6 Possible remainders at the second word for Korean RCs

the predicted Subject Advantage at the head noun derives mainly from entropy differences at the verb immediately preceding it.

This explanation of the Subject Advantage at the head noun accords with that given in Yun, Whitman, and Hale (2010) (YWH). YWH attributed the ORC processing penalty to higher entropy due to additional possible remainders at the verb.[7] YWH observed that in the SRC prefix *N Acc Vtn* the case-marked noun is an argument of the embedded transitive verb as in (9), whereas in the ORC-initial substring *N Nom Vtn* there is an additional possibility that the case-marked noun is in fact a matrix subject, where both the subject and the object of the embedded clause are omitted, as in (10). Since verbal arguments may be freely omitted in Korean when they are recoverable from the context, it is not unreasonable to suppose that this additional possibility, with multiple null elements, indeed plays a role. YWH contains examples of sentences that correspond to the additional structure in (10b) as in (11a), (11b), and (11c). Indeed, adnominal clauses with null elements both in subject and object positions are attested in the corpus as shown in Appendix A.

[7] These "additional" remainders are members of the set-difference between two sets that, due to recursion, are infinite. As Appendix B discusses in further detail, this engenders neither philosophical nor practical difficulty.

(9) SRC-initial substring *N Acc Vtn*

a. [*e* N Acc Vtn]

(10) ORC-initial substring *N Nom Vtn*

a. [N Nom *e* Vtn]

b. N Nom [*e e* Vtn]

(11) Additional possible structures for the ORC-initial substring

a. SRC with Sbj-*pro*

kica ka [*e e* kongkyekhan] uywon ul manassta.
reporter Nom *t pro* attack.Adn senator Acc meet.Decl

'The reporter met the senator who attacked someone.'

b. ORC with Obj-*pro*

kica ka [*e e* kongkyekhan] uywon ul manassta.
reporter Nom *pro t* attack.Adn senator Acc meet.Decl

'The reporter met the senator whom someone attacked.'

c. Noun Complement Clause (NCC) with Sbj- and Obj-*pro*

kica ka [*e e* kongkyekhan] sasil ul alkoissta.
reporter Nom *pro pro* attack.Adn fact Acc know.Decl

'The reporter knows the fact that someone attacked someone.'

The present study confirms this account by tracking all remainders, at each initial substring, as shown in Figure 7. This Figure illustrates that the ORC-initial substring licenses more remainders than does the corresponding SRC-initial substring at the same level of embedding. The remainders ranked 6, 17, and 21 in the (b) panel of Figure 7 correspond to the additional structures (11a)-(11c) that were originally identified in Yun et al. (2010).[8] Part of the ambiguity due to these additional structures is resolved at the next word as the possibility of an NCC

[8] Note that it is possible for the SRC-initial substring to have similar additional derivations if the accusative case marked noun is scrambled over the complex subject modified by an RC or NCC containing both null subject and object. However, the probability of such a construction is quite low. For instance, none of the sentences with adnominal clauses containing two null elements found in the corpus involve scrambling.

is eliminated when the head noun N is heard. This contributes to greater Entropy Reduction for the ORC at the head noun.

Rank	Probability	Remainder	Type
...	...	...	...
5	0.015	*pro* [*t* ***N Acc Vtn***] *N Acc Vtd*	SRC
...	...	...	...
11	0.010	*pro* [*pro* ***N Acc Vtn***] *fact Acc Vtd*	NCC with Sbj-*pro*
...	...	...	...
entropy = 12.905			

(a) SRC-initial substring *N Acc Vtn*

Rank	Probability	Remainder	Type
...	...	...	...
5	0.016	*pro* [***N Nom*** *t* ***Vtn***] *N Acc Vtd*	ORC
6	0.016	***N Nom*** [*pro t* ***Vtn***] *N Acc Vtd*	ORC with Sbj-*pro*
...	...	...	...
17	0.004	***N Nom*** [*t pro* ***Vtn***] *N Acc Vtd*	SRC with Obj-*pro*
...	...	...	...
20	0.003	*pro* [***N Nom*** *pro* ***Vtn***] *fact Acc Vtd*	NCC with Obj-*pro*
21	0.003	***N Nom*** [*pro pro* ***Vtn***] *fact Acc Vtd*	NCC with Sbj- and Obj-*pro*
...	...	...	...
entropy = 14.827			

(b) ORC-initial substring *N Nom Vtn*

Fig. 7 Selected possible remainders at the third word for Korean RCs

6.2 Japanese

Entropy Reduction also predicts the Subject Advantage in Japanese. The word-by-word processing difficulty predictions for Japanese RCs are shown in Figure 8. The pattern here is very similar to the Korean case discussed in subsection 6.1: the Subject Advantage is predicted to show up at the same structural positions as in the Korean examples, namely at the second word (i.e. case marker) and at the fourth word (i.e. head noun). Figure 9 tracks how the most probable structures and their probabilities change as the comprehender encounters new words in Japanese.

The Subject Advantage at the case marker (w_2) is predicted because remainder-set entropy is reduced in the transition from w_1 to w_2 when processing the ORC (ER 1.26 bits), whereas there is no reduction in the case of the SRC. As in Korean, greater processing difficulty for the ORC at w_2 is attributable to the lower entropy

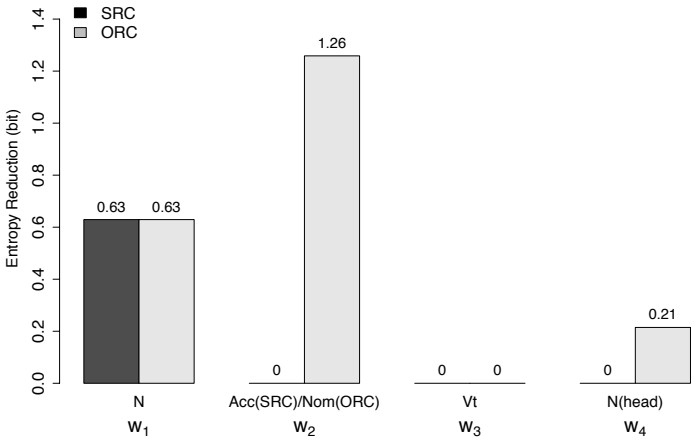


Fig. 8 ER prediction for Japanese RCs

of the ORC-initial substring *N Nom* (4.430 bits), compared to the SRC-initial substring *N Acc* (5.773 bits) since the entropy at w_1 is the same for the SRC and the ORC (5.688 bits). Although the difference in entropy at w_2 is not immediately obvious from the very top-ranked parses as it was in the case of Korean, the source of the difference turns out to be similar with the Korean case if we continue by examining some lower-order derivations. Figure 10 shows all analyses with a probability over 0.001 for both SRC and ORC initial substrings at the case marker. The smaller cardinality (31 vs 42) of nontrivial derivations in the ORC already suggests that more work has been done, resulting in a more highly-organized parser state. Further inspection of Figure 10 shows that the remainders of *N Nom* tend to concentrate their probability mass on interpretations of that string as a subject of a simple sentence. On the other hand, the remainders of *N Acc* exhibit a broader spread over more complex derivations involving more levels of embedding. The difference between these probability distributions creates the Subject Advantage at the case marker, in much the same way as in Korean. Korean and Japanese seem to differ here only in degree.

The sources for the Subject Advantage at the head noun (w_4) are, however, not exactly parallel with Korean, due to a distinctive characteristic of Japanese. Note that in Japanese, adnominal and declarative forms are the same for verbs.

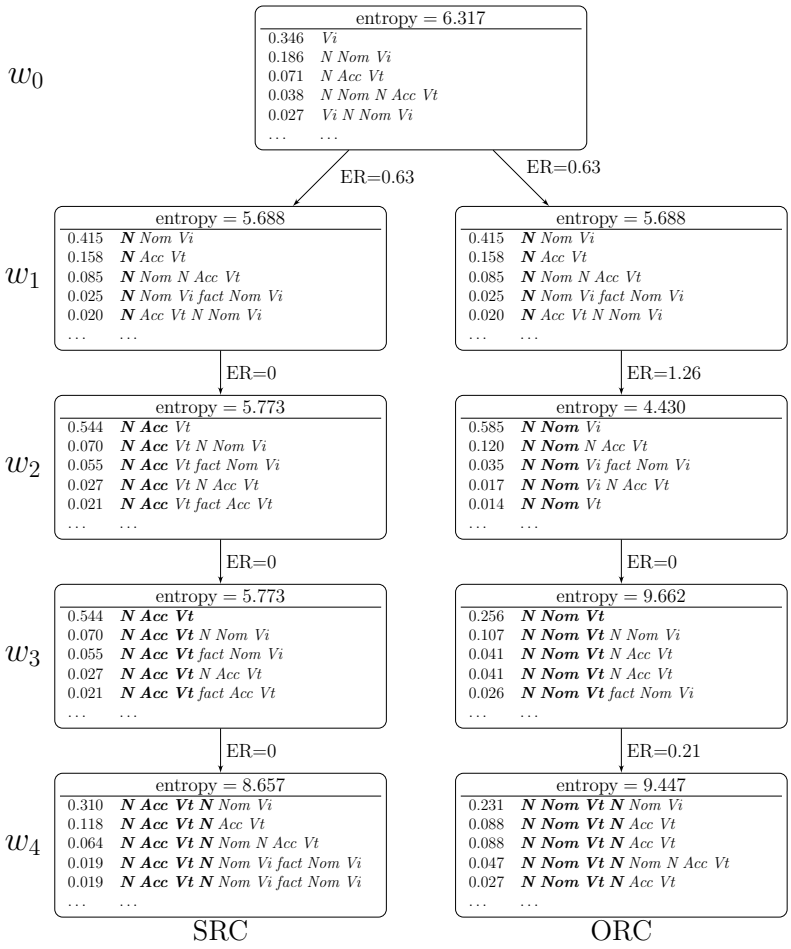


Fig. 9 Likely derivations and their conditional probabilities in Japanese RCs

Thus when the third word (i.e. the verb) is heard, this sentence-initial substring is ambiguous between adnominal clauses (i.e. relative clause or noun complement clauses) and declarative clauses. Figure 11 shows that both the SRC and ORC at w_3 are most likely to be interpreted as declarative sentences at this point. However, the distribution of the remainder set is more concentrated around a simple declarative analysis in the case of the SRC-initial substring (0.544), compared to the ORC-initial substring (0.256). This difference in the distribution of the remainder sets reflects the asymmetrical distribution of subject and object *pro*s in Japanese: subject *pro* is abundant, while object *pro* is less so. Appendix A details

Rank	Prob	Remainder
1.	0.544	*pro* ***N Acc*** *Vt*
2.	0.070	[*t* ***N Acc*** *Vt*] *N Nom Vi*
3.	0.055	[*pro* ***N Acc*** *Vt*] *fact Nom Vi*
4.	0.027	*pro* [*t* ***N Acc*** *Vt*] *N Acc Vt*
5.	0.021	*pro* [*pro* ***N Acc*** *Vt*] *fact Acc Vt*
6.	0.014	[*t* ***N Acc*** *Vt*] *N Nom N Acc Vt*
7.	0.011	[*pro* ***N Acc*** *Vt*] *fact Nom N Acc Vt*
8.	0.010	[*t* ***N Acc*** *Vt*] [*t Vi*] *N Nom Vi*
9.	0.008	[*t* ***N Acc*** *Vt*] [*pro Vi*] *fact Nom Vi*
10.	0.004	[[*t* ***N Acc*** *Vt*] *N Nom Vi*] *fact Nom Vi*
11.	0.004	[*t* ***N Acc*** *Vt*] [*N Nom Vi*] *fact Nom Vi*
12.	0.004	*pro* [*t* ***N Acc*** *Vt*] [*t Vi*] *N Acc Vt*
13.	0.003	[*t* ***N Acc*** *Vt*] [*t N Acc Vt*] *N Nom Vi*
14.	0.003	[*t* [*t* ***N Acc*** *Vt*] *N Acc Vt*] *N Nom Vi*
15.	0.003	[[*pro* ***N Acc*** *Vt*] *fact Nom Vi*] *fact Nom Vi*
16.	0.003	*pro* [*t* ***N Acc*** *Vt*] [*pro Vi*] *fact Acc Vt*
17.	0.003	[*t* ***N Acc*** *Vt*] [*pro N Acc Vt*] *fact Nom Vi*
18.	0.003	[*t* [*pro* ***N Acc*** *Vt*] *fact Acc Vt*] *N Nom Vi*
19.	0.003	[*pro* [*t* ***N Acc*** *Vt*] *N Acc Vt*] *fact Nom Vi*
20.	0.002	[*pro* [*pro* ***N Acc*** *Vt*] *fact Acc Vt*] *fact Nom Vi*
21.	0.002	[*t* ***N Acc*** *Vt*] *N Nom* [*t Vi*] *N Acc Vt*
22.	0.002	[*t* ***N Acc*** *Vt*] [*t Vi*] *N Nom N Acc Vt*
23.	0.002	[*t* ***N Acc*** *Vt*] *N Nom pro Vt*
24.	0.002	[*t* ***N Acc*** *Vt*] *N Nom* [*pro Vi*] *fact Acc Vt*
25.	0.002	[*t* ***N Acc*** *Vt*] [*pro Vi*] *fact Nom N Acc Vt*
26.	0.002	[*pro* ***N Acc*** *Vt*] *fact Nom* [*t Vi*] *N Acc Vt*
27.	0.002	*pro* [[*t* ***N Acc*** *Vt*] *N Nom Vi*] *fact Acc Vt*
28.	0.002	*pro* [*t* ***N Acc*** *Vt*] [*N Nom Vi*] *fact Acc Vt*
29.	0.001	[[*t* ***N Acc*** *Vt*] *N Nom N Acc Vt*] *fact Nom Vi*
30.	0.001	[*t* ***N Acc*** *Vt*] [*N Nom N Acc Vt*] *fact Nom Vi*
31.	0.001	[*t* ***N Acc*** *Vt*] [*t Vi*] [*t Vi*] *N Nom Vi*
32.	0.001	*pro* [*t* ***N Acc*** *Vt*] [*t N Acc Vt*] *N Acc Vt*
33.	0.001	*pro* [*t* [*t* ***N Acc*** *Vt*] *N Acc Vt*] *N Acc Vt*
34.	0.001	[*pro* ***N Acc*** *Vt*] *fact Nom pro Vt*
35.	0.001	[*t* ***N Acc*** *Vt*] [*pro t Vt*] *N Nom Vi*
36.	0.001	[*pro* ***N Acc*** *Vt*] *fact Nom* [*pro Vi*] *fact Acc Vt*
37.	0.001	*pro* [[*pro* ***N Acc*** *Vt*] *fact Nom Vi*] *fact Acc Vt*
38.	0.001	[[*pro* ***N Acc*** *Vt*] *fact Nom N Acc Vt*] *fact Nom Vi*
39.	0.001	[*t* ***N Acc*** *Vt*] [*t Vi*] [*pro Vi*] *fact Nom Vi*
40.	0.001	*pro* [*t* ***N Acc*** *Vt*] [*pro N Acc Vt*] *fact Acc Vt*
41.	0.001	*pro* [*t* [*pro* ***N Acc*** *Vt*] *fact Acc Vt*] *N Acc Vt*
42.	0.001	*pro* [*pro* [*t* ***N Acc*** *Vt*] *N Acc Vt*] *fact Acc Vt*
entropy = 5.773		

(a) SRC-initial substring *N Acc*

Rank	Prob	Remainder
1.	0.585	***N Nom*** *Vi*
2.	0.120	***N Nom*** *N Acc Vt*
3.	0.035	[***N Nom*** *Vi*] *fact Nom Vi*
4.	0.017	***N Nom*** [*t Vi*] *N Acc Vt*
5.	0.014	***N Nom*** *pro Vt*
6.	0.013	*pro* [***N Nom*** *Vi*] *fact Acc Vt*
7.	0.013	***N Nom*** [*pro Vi*] *fact Acc Vt*
8.	0.012	[***N Nom*** *N Acc Vt*] *fact Nom Vi*
9.	0.007	[***N Nom*** *Vi*] *fact Nom N Acc Vt*
10.	0.007	***N Nom*** [*N Nom Vi*] *fact Acc Vt*
11.	0.006	***N Nom*** [*t N Acc Vt*] *N Acc Vt*
12.	0.006	[***N Nom*** *t Vt*] *N Nom Vi*
13.	0.005	*pro* [***N Nom*** *N Acc Vt*] *fact Acc Vt*
14.	0.005	***N Nom*** [*pro N Acc Vt*] *fact Acc Vt*
15.	0.002	[***N Nom*** *N Acc Vt*] *fact Nom N Acc Vt*
16.	0.002	***N Nom*** [*N Nom N Acc Vt*] *fact Acc Vt*
17.	0.002	***N Nom*** [*t Vi*] [*t Vi*] *N Acc Vt*
18.	0.002	*pro* [***N Nom*** *t Vt*] *N Acc Vt*
19.	0.002	***N Nom*** [*pro t Vt*] *N Acc Vt*
20.	0.002	[[***N Nom*** *Vi*] *fact Nom Vi*] *fact Nom Vi*
21.	0.002	***N Nom*** [*t Vi*] [*pro Vi*] *fact Acc Vt*
22.	0.002	[***N Nom*** [*t Vi*] *N Acc Vt*] *fact Nom Vi*
23.	0.002	[*t* [***N Nom*** *Vi*] *fact Acc Vt*] *N Nom Vi*
24.	0.001	[***N Nom*** *pro Vt*] *fact Nom Vi*
25.	0.001	[*pro* [***N Nom*** *Vi*] *fact Acc Vt*] *fact Nom Vi*
26.	0.001	[***N Nom*** [*pro Vi*] *fact Acc Vt*] *fact Nom Vi*
27.	0.001	[***N Nom*** *t Vt*] *N Nom N Acc Vt*
28.	0.001	***N Nom*** [*N Nom t Vt*] *N Acc Vt*
29.	0.001	[***N Nom*** *Vi*] *fact Nom* [*t Vi*] *N Acc Vt*
30.	0.001	***N Nom*** [*t Vi*] [*N Nom Vi*] *fact Acc Vt*
31.	0.001	***N Nom*** [[*t Vi*] *N Nom Vi*] *fact Acc Vt*
entropy = 4.430		

(b) ORC-initial substring *N Nom*

Fig. 10 Expanded derivation lists ($p > 0.001$) at the second words in Japanese RCs

how this distributional difference manifests itself in Japanese corpora. If a comprehender detects a missing subject in the sentence, the empty category is likely to be *pro*, whereas a missing direct object is rather ambiguous between an RC gap and *pro*. Thus, a hearer can be more certain about the rest of the sentence having heard the SRC-initial substring *N Acc Vt*, which lacks an overt subject, compared to the ORC-initial substring *N Nom Vt*, which lacks an overt direct object. The lower entropy at w_3 for the SRC-initial substring (5.773 bits) than for the ORC-initial substring (9.662 bits) quantifies this claim.[9] The entropy at the SRC-initial substring at w_3 is in fact so low that no Reduction happens at all in the

[9] As in Korean, the additional remainders for the ORC-initial substring also contribute to its high entropy. However, the influence of ambiguous verb forms seems much stronger in processing Japanese RCs, as indicated by the highly asymmetric probability distributions of the SRC and ORC remainder-sets at the verb.

transition from w_3 to w_4 in the SRC. This contributes to the Subject Advantage at the head noun (w_4) in Japanese.

Probability	Remainder	Type
0.544	*pro* ***N Acc Vt***	simplex SOV with Sbj-*pro*
0.070	[*t* ***N Acc Vt***] *N Nom Vi*	SRC
0.055	[*pro* ***N Acc Vt***] *fact Nom Vi*	NCC with Sbj-*pro*
0.027	*pro* [*t* ***N Acc Vt***] *N Acc Vt*	SRC
0.021	*pro* [*pro* ***N Acc Vt***] *fact Acc Vt*	NCC with Sbj-*pro*
...	...	...
Entropy = 5.773		

(a) SRC prefix *N Acc Vt*

Probability	Remainder	Type
0.256	[***N Nom*** *pro* ***Vt***]	Simplex SOV with Obj-*pro*
0.107	[***N Nom*** *t* ***Vt***] *N Nom Vi*	ORC
0.041	*pro* [***N Nom*** *t* ***Vt***] *N Acc Vt*	ORC
0.041	***N Nom*** [*pro t* ***Vt***] *N Acc Vt*	ORC with Sbj-*pro*
0.026	[***N Nom*** *pro* ***Vt***] *fact Nom Vi*	NCC with Obj-pro
...	...	...
Entropy = 9.662		

(b) ORC prefix *N Nom Vt*

Fig. 11 Possible remainders at the third word for Japanese RCs

6.3 Chinese

The true processing difficulty profile of Chinese RCs has been tough to pin down. One of the factors that confounds research on SRCs and ORCs in this language is the presence of temporary ambiguities. A variety of syntactic alternatives are compatible with the initial substrings of these constructions. For example, the SRC-initial substring (Vt N ...) can also be understood as a *pro*-dropped matrix clause. Given that subject *pro*-drop in Chinese is extremely frequent,[10] this possibility must be taken into account. ORCs, too, are plagued by troublesome temporary ambiguities. For instance, the ORC-initial substring (N Vt ...) may also continue as a matrix clause. If readers have a strong expectation for this

[10] In our Chinese Treebank search (see Appendix A.3 for more results), we find that there are even more omitted arguments (6385) than bare nouns (3830) in subject position.

structure, they may not realize they have read an ORC. Moreover, when an ORC modifies the matrix object, the initial substring (N Vt [N Vt ...) can trigger a garden-path effect of reanalysis from a matrix clause to an RC (Lin and Bever, 2006, 2011).

In order to test whether the Subject Advantage indeed exists in Chinese RCs, Jäger et al. (submitted) used an experimental design where RC-initial substrings are disambiguated from matrix clauses. The disambiguation is accomplished using extra words that help guide readers towards some RC interpretation while still leaving the specific gap site unspecified.

(12) a. "Disambiguated" Chinese SRC

na ge zuotian [e_i yaoqing fuhao de] guanyuan$_i$
that.Dem Cl yesterday.Time [*t* invite.Vt tycoon.N DE] official.N
'the official who invited the tycoon yesterday'

b. "Disambiguated" Chinese ORC

na ge zuotian [fuhao yaoqing e_i de] guanyuan$_i$
that.Dem Cl yesterday.Time [tycoon.N invite.Vt *t* DE] official.N
'the official who the tycoon invited yesterday'

In example (12), the sentence-initial demonstrative-classifier combination "na-ge" encourages readers to expect a noun phrase. However, the following word is a temporal phrase "yesterday" which has to be attached to a verb phrase. This design therefore leads the reader to only foresee an upcoming RC-modified noun phrase by ruling out the *pro*-drop analysis. Jäger et al. (submitted) used these "disambiguated" RC stimuli in both self-paced reading and eye-tracking experiments. They reported that SRCs are consistently read faster than ORCs in the RC region (*Vt N* or *N Vt*, respectively) and at the head noun. A Subject Advantage was also found after the head noun, potentially a spillover effect from previous regions.

Entropy Reduction derives a Subject Advantage in Chinese RCs as shown in Figure 12. ORCs are correctly categorized as harder to process, in both the RC and head noun regions.

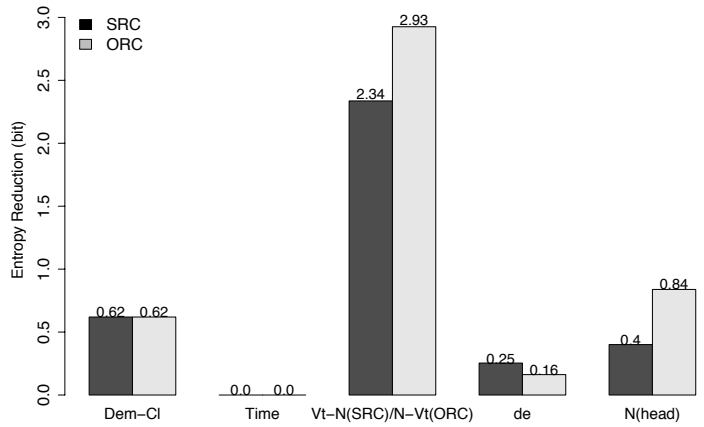


Fig. 12 ER predictions of "disambiguated" Chinese RCs

Figure 13 illustrates how the Subject Advantage is predicted in Chinese RCs region-by-region. Because the word order of SRCs (*Vt N*) is not the same as that of ORCs (*N Vt*) in Chinese, we collapse the two words in the RC region. After processing this RC region, entropy reduces by 2.34 bits in the SRC case and 2.93 bits in the ORC case. Looking at the boxes in the "RC region" row, we focus on the two initial substrings *Dem Cl Time Vt N* and *Dem Cl Time N Vt*.

The greater Entropy Reduction in the ORC case as compared to the SRC case quantifies the intuitive idea that more comprehension work is called for in the RC region of one item as compared to the other. This numerical contrast reflects the fact that, by the end of the RC region, the ORC parser state is more organized than the corresponding SRC state. Table 5 quantifies this degree of organization by counting the remainders at various probability thresholds. While both initial substrings start from the same uncertainty level (7.448 bits), the ORC-initial substring goes farther in reducing this ambiguity. In other words, more work has been done.

Earlier modeling work on Chinese RCs using Surprisal did not derive the Subject Advantage at the RC head noun (Z. Chen, Grove, and Hale, 2012); however, the present work with Entropy Reduction does this. Examining the pre-head syntactic remainders in Figure 14, the grammar defines contrasting probabilities for

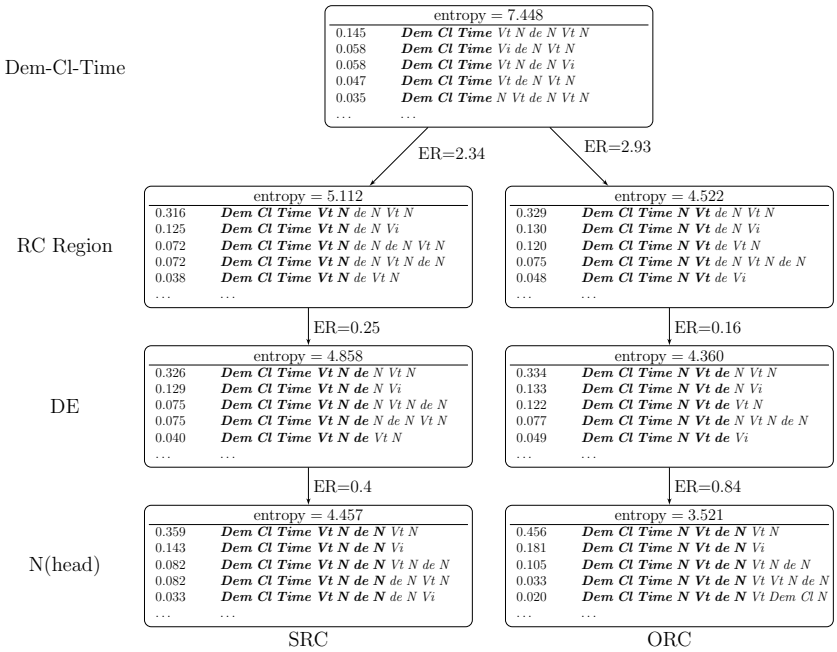


Fig. 13 Derivations and their conditional probabilities in "disambiguated" Chinese RCs

Probability	Type	No. of Remainders	Total Probability	Uncertainty
> 0.01	SRC	10	0.722	High
	ORC	8	0.768	Low
> 0.001	SRC	64	0.881	High
	ORC	54	0.908	Low

Table 5 More peaked probability distribution of remainders leads to lower uncertainty level in ORCs. Starting from the same entropy level (first row of Figure 13) but arriving at a lower "destination" level (second row of Figure 13) in just the ORC yields the prediction of greater processing effort via the Entropy Reduction hypothesis.

headlessness in SRC vs ORC contexts. In ORCs, as shown in the bottom left box, there is at least a 20.8% chance that the initial substring will continue as a headless RC. On the other hand, it is less likely (5.6%) that an SRC-initial substring will be headless (top left box). The information-value of the headlessness question itself is different across the two constructions, and this contributes to the prediction of a Subject Advantage at the head word in Chinese.

SRC

Prob	Remainder
0.326	*Dem Cl Time Vt N de N Vt N*
0.129	*Dem Cl Time Vt N de N Vi*
0.075	*Dem Cl Time Vt N de N Vt N de N*
0.075	*Dem Cl Time Vt N de N de N Vt N*
0.040	***Dem Cl Time Vt N de Vt N***
0.030	*Dem Cl Time Vt N de N de N Vi*
0.023	*Dem Cl Time Vt N de N Vt Vt N de N*
0.017	*Dem Cl Time Vt N de N de N Vt N de N*
0.016	***Dem Cl Time Vt N de Vi***
0.014	*Dem Cl Time Vt N de N Vt Dem Cl N*
...	...
entropy = 4.858	

ER=0.40

Prob	Remainder
0.359	*Dem Cl Time Vt N de N Vt N*
0.143	*Dem Cl Time Vt N de N Vi*
0.082	*Dem Cl Time Vt N de N Vt N de N*
0.082	*Dem Cl Time Vt N de N de N Vt N*
0.033	*Dem Cl Time Vt N de N de N Vi*
0.026	*Dem Cl Time Vt N de N Vt Vt N de N*
0.019	*Dem Cl Time Vt N de N de N Vt N de N*
0.016	*Dem Cl Time Vt N de N Vt Dem Cl N*
0.010	*Dem Cl Time Vt N de N Vt Vi de N*
0.010	*Dem Cl Time Vt N de N de Vt N*
...	...
entropy = 4.457	

ORC

Prob	Remainder
0.334	*Dem Cl Time N Vt de N Vt N*
0.132	*Dem Cl Time N Vt de N Vi*
0.122	***Dem Cl Time N Vt de Vt N***
0.077	*Dem Cl Time N Vt de N Vt N de N*
0.049	***Dem Cl Time N Vt de Vi***
0.028	***Dem Cl Time N Vt de Vt N de N***
0.024	*Dem Cl Time N Vt de N Vt Vt N de N*
0.015	*Dem Cl Time N Vt de N Vt Dem Cl N*
0.010	*Dem Cl Time N Vt de N Vt Vi de N*
0.009	***Dem Cl Time N Vt de Vt Vt N de N***
...	...
entropy = 4.360	

ER=0.84

Prob	Remainder
0.456	*Dem Cl Time N Vt de N Vt N*
0.181	*Dem Cl Time N Vt de N Vi*
0.105	*Dem Cl Time N Vt de N Vt N de N*
0.033	*Dem Cl Time N Vt de N Vt Vt N de N*
0.020	*Dem Cl Time N Vt de N Vt Dem Cl N*
0.013	*Dem Cl Time N Vt de N Vt Vi de N*
0.012	*Dem Cl Time N Vt de N Vt de N Vt N*
0.011	*Dem Cl Time N Vt de N Vt Vt de N*
0.008	*Dem Cl Time N Vt de N Vt de N Vt N*
0.008	*Dem Cl Time N Vt de N Vt de N Vt N*
...	...
entropy = 3.521	

Fig. 14 ER at the head noun in Chinese RCs; detail of last two rows in Figure 13. Here, boldface highlights expectations that the RC will be headless.

7 Conclusion

Processing asymmetries across relative clause types can be understood as differential degrees of confusion about sentence structure. On the one hand, this idea has a certain intuitive obviousness about it: a harder information-processing problem ought to be associated with greater observable processing difficulty. But it is not a foregone conclusion. It could have been the case that the sentence structures motivated by leading syntactic analyses do not derive these "harder" vs "easier" information processing problems in just the cases where human readers seem to have trouble. But in fact, as section 6 details, they do.

By suggesting that human listeners are information-processors subject to information theory, this methodology revives the approach of psychologists like Hick (1952), Attneave (1959) and Garner (1962). What differs in the present application of information theory is the use of generative grammars to specify the sentence-structural expectations that hearers have. The conditional probabilities in diagrams like Figure 5 are calculated over *grammatical* alternatives, not just single successor words. In this regard our method integrates both information-processing psychology and generative grammar.

This integration articulates the view that all humans are bound by the same kinds of processing limitations. Where the distribution on grammatical alternatives goes from being flatter to being more peaked, we should expect slower reading times. Such a limitation may perhaps reflect something deep about cognition, something that constrains both humans and machines. However, the precise implications of this limit depend crucially on the grammar of particular languages.

Acknowledgements

The research reported in this article was supported by NSF CAREER award number 0741666. JY and JBW were partially supported by a lab grant from the Academy of Korean Studies.

A Attestation counts

A.1 Korean

Korean Treebank 2.0 (Han et al., 2002) was used to obtain an estimation of the frequencies of the relevant structures in Korean. The texts of the corpus are a selection of Korean Press Agency news articles in 2000, consisting of 5,010 sentences and 132,040 words. The weight parameters for the MG presented in example 6 on page 16 were derived from the attestation counts given below in Table 6.

Clause Type	Subject	Object	Verb	Count
Matrix Clause	overt	overt	transitive	24
	overt	pro	transitive	2
	pro	overt	transitive	31
	pro	pro	transitive	0
	overt	-	intransitive	436
	pro	-	intransitive	30
Complement Clause	overt	overt	transitive	92
	overt	pro	transitive	5
	pro	overt	transitive	401
	pro	pro	transitive	12
	overt	-	intransitive	282
	pro	-	intransitive	110
Relative Clause (SRC)	gap	overt	transitive	467
	gap	pro	transitive	4
	gap	-	intransitive	559
Relative Clause (ORC)	overt	gap	transitive	120
	pro	gap	transitive	10

Table 6 Korean attestation counts

A.2 Japanese

In the case of Japanese, Kyoto Corpus 4.0 (Kurohashi and Nagao, 2003) was used. The texts of the corpus are a selection of Mainichi Newspaper articles from 1995, consisting of 5,447 sentences and 161,028 words. Although the Kyoto Corpus is not a treebank, it does provide a rich set of part-of-speech tags that we used to carry out a corpus study analogous to the Korean study discussed above. In particular, all predicates are annotated with information about their arguments, including the location of the argument and its syntactic type (such as nominative or accusative). Although Japanese RCs are prenominal, the canonical word order places the verb at the end of the clause, after all of its arguments. Consequently, if a verb is followed by its nominative or accusative argument, the clause that ends with the verb is an SRC or ORC, respectively. If a verb comes after all its arguments but still precedes some other noun, then the clause that ends with the verb is a noun complement clause. The frequencies of the same parameters as in Korean were derived from the following counts.

Clause Type	Subject	Object	Verb	Count
Matrix Clause	overt	overt	transitive	125
	overt	pro	transitive	7
	pro	overt	transitive	329
	pro	pro	transitive	32
	overt	-	intransitive	691
	pro	-	intransitive	487
Complement Clause	overt	overt	transitive	149
	overt	pro	transitive	4
	pro	overt	transitive	323
	pro	pro	transitive	25
	overt	-	intransitive	463
	pro	-	intransitive	200
Relative Clause (SRC)	gap	overt	transitive	537
	gap	pro	transitive	20
	gap	-	intransitive	854
Relative Clause (ORC)	overt	gap	transitive	116
	pro	gap	transitive	102

Table 7 Japanese attestation counts

A.3 Chinese

We obtain attestation counts from Chinese Treebank 7 (Xue et al., 2005) which contains 51,447 fully parsed sentences or 1,196,329 words. These yield the weights shown below in Table 8. Note that the "disambiguated" RCs shown in example 12 on page 32 motivate a somewhat richer set of choice points in the formal grammar fragment, which obligates us to estimate weights for a longer list of parameters than in Korean or Japanese.

noun with a demonstrative modifier	2916
complex NP with a demonstrative modifier	345
noun in argument position	8133
complex NP in argument position	2316
possessive phrase in argument position	1866
headful SRC	2281
headless SRC	280
headful ORC	830
headless ORC	304
noun in subject position	3830
noun with a demonstrative modifier in subject position	167
pro in subject position	6385
noun in object position	3766
noun with a demonstrative modifier in object position	123
pro in object position	2
subject *pro* with transitive verb	5054
subject *pro* with intransitive verb	1331
subject NP with transitive verb	17250
subject NP with intransitive verb	4377
noun as ORC subject	185
noun with a demonstrative modifier as ORC subject	12
pro as ORC subject	162
matrix modified by temporal adjunct	343
matrix not modified by temporal adjunct	16852
SRC not modified by temporal adjunct	2532
ORC not modified by temporal adjunct	1124
RC modified by temporal adjunct	39
relative clause	3695
complement clause	852

Table 8 Chinese weights

B A note on the infinity of possible remainders

References to dozens of possible sentence-remainders in, for example, Table 5 or Figure 7, might prompt the question of whether our account assumes some degree of parallelism in the parsing process. It is true that, for example, (the probability of) the 100^{th}-best possible remainder plays a role in determining the predictions of the Entropy Reduction Hypothesis, but this does not entail any algorithmic claim that the comprehender in fact proceeds by considering each of the top 100 derivations one by one.

Instead of a processing algorithm, the ER complexity metric rather models a comprehender's intermediate mental state at a particular position in a sentence using a *grammar*. Section 5 dubs this sort of object an "intersection grammar". Such a grammar is a finite object that characterizes the set of possible sentence-remainders, which is in many cases an infinite set. This characterization is precisely analogous to the way that generative grammars are ordinarily used to characterize presumably infinite sets of well-formed *un*-remaindered, full sentences.[11]

The important idea is that when we report, for example, an entropy value of 5.773 after encountering the initial substring *N Acc* in Figure 9, this value is a property *of the intersection grammar* that finitely characterizes the comprehender's mental state at that point; we can ask the question of what this entropy value is in much the same way that we can ask, say, how many rules are in the grammar, or what the greatest number of symbols on the right hand side of a single rule is. (These are other conceivable, though perhaps not so well-motivated, linking hypotheses one might consider using to connect intersection grammars to behavioral predictions.) In order to gain some understanding of the content of a remainder grammar, it is useful to consider some of the infinitely many derivations that it licenses, and this is the purpose of derivation lists that appear throughout section 6. These lists help to understand the implications of the remainder grammars, but it is the remainder grammars' uncertainty properties, and not the lists themselves that we offer as a cognitive account of comprehenders' intermediate mental states.

To illustrate this point, consider the grammar in Figure 15. This finite object derives infinitely many sentences such as the following (where the brackets are added just for readability).

(13) a. The fact that [John met Mary] shows that [the cat jumped]

b. The fact that [the claim that [the cat jumped] was rejected] suggests that [the report that [John met Mary] was true]

[11] In the special case where we consider the comprehender's mental state at the beginning of a sentence, i.e. where the initial substring encountered so far is the empty string, these two ideas reduce to the same thing: the possible remainders consistent with this empty initial substring are precisely the well-formed sentences of the language.

c. ...

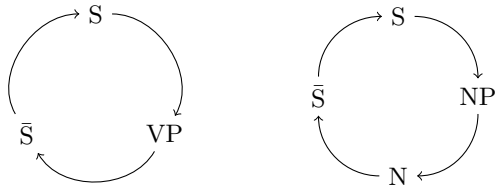
This grammar licenses an infinity of derivations because it contains "loops". For example, an S can contain a VP, which can contain an $\bar{\text{S}}$, which can contain an S, as illustrated in Figure 16(a). There is also a second, longer, loop, consisting of four nonterminals, shown in Figure 16(b).

Consider now a comprehender using this grammar to parse a sentence, at the point where only the initial substring *The fact that John met Mary shows* has been encountered. Our methodology characterizes this comprehender's mental state at this point by an intersection grammar, a modified version of the grammar in Figure 15 that derives only sentences consistent with this initial substring. Clearly there are an infinite number of such sentences, since they have the form *The fact that John met Mary shows that S*; hence the intersection grammar must necessarily contain at least one loop. Note that this is true even though there are an infinite number of other sentences that are no longer candidates because they are not consistent with the heard initial substring, but which are derivable using the grammar in Figure 15, since the main clause subject can contain unboundedly many embeddings. More specifically, then, the remainder grammar will have in effect the same two loops shown in Figure 16, although the ways these loops can be reached are more restricted than is the case for the grammar in Figure 15.

The import of this example for present purposes is this: just as the infinite list of sentences in this imagined language can be characterized by the finite grammar in Figure 15, and hence by a certain finite mental state, so can the infinite list of sentences consistent with the relevant initial substring. In order to gain some intuitive understanding of what is characterized by such a finite mental state, it is often useful to "unroll" some of the licensed derivations for inspection; this is what we are doing when we look at (13) to gain some understanding of the grammar in Figure 15, and the larger lists in section 6 were used to similarly gain some understanding of certain intersection grammars. These lists themselves, however, play no role in the theory we propose. Just as a competent speaker is usually hypothesized to bear a certain mental relationship to a grammar such as the one in Figure 15, rather than any particular elements of or subsets of lists like (13), on our view a comprehender bears a certain mental relationship to a intersection grammar (which, in turn, has an entropy), rather than any particular elements of or subsets of lists like the ones presented in preceding subsections.

S	→	NP VP
NP	→	the N
N	→	cat | dog | ... | fact $\bar{\text{S}}$ | claim $\bar{\text{S}}$
VP	→	jumped | is true | was rejected | ... | shows $\bar{\text{S}}$ | suggests $\bar{\text{S}}$
$\bar{\text{S}}$	→	that S

Fig. 15 A simple grammar for an imagined language



(a) One loop (b) Another loop

Fig. 16 The two loops present in the grammar in Figure 15

C Entropy Reduction and Information Theory

The reduction in entropy of grammatical derivations brought about by observing (or lengthening) an initial substring represents a particular way of characterizing information gain. This particular way is written out as the difference of two entropies in 14 below. In this definition, X is a random variable over derivations. $Y = y$ denotes the outcome of a related random variable such as an initial substring of the yield.

$$I(X; y) = H(X) - H(X|y) \tag{14}$$

Blachman (1968) compares the information measure in 14 to an alternative measure given below in 15.

$$J(X; y) = \sum_x p(x|y) \log \left(\frac{p(x|y)}{p(x)} \right) \tag{15}$$

The second measure, J, leads to surprisal in the sense of Hale (2001) via reasoning analogous to that presented in section 2.1 of Levy (2008). The first measure, I, leads to Entropy Reduction. As Blachman points out, both notions of information gain have as their expectation the *mutual information* of X and Y. However, he goes on to show that they are not equivalent. For instance, I can be negative whereas J cannot. On the other hand, I is additive while J is not.

What this shows is that there is no one "official" way to apply information theory to human language processing. These definitions are theoretical postulations that are useful to the extent that they lead to greater insight into the phenomena under study.

References

Attneave, Fred (1959). *Applications of Information Theory to Psychology: A summary of basic concepts, methods and results*. Holt, Rinehart and Winston.

Bar-Hillel, Yehoshua, Micha Perles, and Eliyahu Shamir (1964). On formal properties of simple phrase structure grammars. In *Language and Information: Selected Essays on their Theory and Application*, Chapter 9, pp. 116–150. Reading, Massachusetts: Addison-Wesley.

Berwick, Robert and Samuel Epstein (1995). On the Convergence of 'Minimalist' Syntax and Categorial Grammar. In A. Nijholt (Ed.), *Algebraic methods in language processing : proceedings of the tenth Twente Workshop on Language Technology joint with first AMAST Workshop on Language Processing*.

Bever, Thomas (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language*, pp. 279–360. John Wiley.

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

Brame, Michael (1967). A new analysis of the relative clause: evidence for an interpretive theory. unpublished manuscript.

Chen, Baoguo, Aihua Ning, Hongyan Bi, and Susan Dunlap (2008). Chinese subject-relative clauses are more difficult to process than the object-relative clauses. *Acta Psychologica 129*(1), 61–65.

Chen, Zhong (2014). *Animacy in sentence processing across languages: an information-theoretical perspective*. Ph. D. thesis, Cornell University.

Chen, Zhong, Kyle Grove, and John Hale (2012). Structural expectations in Chinese relative clause comprehension. In J. Choi, E. A. Hogue, J. Punske, D. Tat, J. Schertz, and A. Trueman (Eds.), *Proceedings of the 29th West Coast Conference on Formal Linguistics (WCCFL-29)*, Somerville, MA, pp. 29–37. Cascadilla Proceedings Project.

Cherry, Colin (1961). *On human communication: a review, a survey, and a criticism* (2. printing ed.). New York: Science Ed.

Chomsky, Noam (1993). A minimalist program for linguistic theory. In *The view from building 20: essays in linguistics in honor of Sylvain Bromberger*, Volume 24 of *Current studies in linguistics*, pp. 1–52. Cambridge, Mass.: MIT Press.

Chomsky, Noam (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Frank, Stefan (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science 5*(3), 475–494.

Frauenfelder, Uli, Juan Segui, and Jacques Mehler (1980). Monitoring around the relative clause. *Journal of Verbal Learning and Verbal Behavior 19*, 328–337.

Frazier, Lyn (1987). Syntactic processing: Evidence from Dutch. *Natural Language & Linguistic Theory 5*(4), 519–559.

Garner, Wendell (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.

Gibson, Edward (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, and W. O'Neil (Eds.), *Image, Language, brain: Papers from the First Mind Articulation Project Symposium*. Cambridge, MA: MIT Press.

Gibson, Edward and Hsiao-Hung Iris Wu (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes 28*(1-2), 125–155.

Grenander, Ulf (1967). Syntax-controlled probabilities. Technical report, Brown University Division of Applied Mathematics, Providence, RI.

Guillaumin, Matthieu (2005). Conversations between mildly context-sensitive grammars. Internship report, Ecole Normale Superieure and UCLA.

Hale, John (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2^{nd} NAACL*, Pittsburgh, PA.

Hale, John (2003, March). The information conveyed by words in sentences. *Journal of Psycholinguistic Research 32*(2), 101–123.

Hale, John (2006). Uncertainty about the rest of the sentence. *Cognitive Science 30*(4), 643—672.

Han, Chung-hye, Na-Rae Han, Eon-Suk Ko, and Martha Palmer (2002). Development and Evaluation of a Korean Treebank and its Application to NLP. *Language and Information 6*(1), 123–138.

Han, Chung-hye and Jong-Bok Kim (2004). Are there "double relative clauses" in Korean? *Linguistic Inquiry 35*(2), 315–337.

Harkema, Henk (2001). *Parsing minimalist grammars*. Ph. D. thesis, University of California, Los Angeles.

Harris, Theodore (1963). *The Theory of Branching Processes*. Springer-Verlag.

Hawkins, John (2004). *Efficiency and Complexity in Grammars*. Oxford University Press.

Hick, William (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology 4*(1), 11–26.

Hirose, Yuki (2009). Processing relative clauses in Japanese: coping with multiple ambiguities. In *The Handbook of East Asian Psycholinguistics*, Volume II: Japanese, Chapter 35, pp. 264–269. Cambridge: Cambridge University Press.

Hoshi, Koji (1995). *Structural and interpretive aspects of head-internal and head-external relative Structural and interpretive aspects of head-internal and head-external relative clauses*. Ph. D. thesis, University of Rochester.

Hsiao, Franny and Edward Gibson (2003). Processing relative clauses in Chinese. *Cognition 90*, 3–27.

Hsiao, Yaling, Jinman Li, and Maryellen MacDonald (2014). Ambiguity affects Mandarin relative clause processing. In *The 27th Annual CUNY Conference on Human Sentence Processing*, Columbus, OH. The Ohio State University.

Huang, C.-T. James, Yen-Hui Audrey Li, and Yafei Li (2009). *The Syntax of Chinese*. Cambridge University Press.

Ishii, Yasuo (1991). *Operators and empty categories in Japanese*. Ph. D. thesis, University of Connecticut.

Ishizuka, Tomoko (2005). Processing relative clauses in Japanese. In R. Okabe and K. Nielsen (Eds.), *Papers in Psycholinguistics 2*, Volume 13 of *UCLA Working Papers in Linguistics*, pp. 135–157. Los Angeles: UCLA Linguistics Department.

Ishizuka, Tomoko, Kentaro Nakatani, and Edward Gibson (2003). Relative clause extraction complexity in Japanese. Poster presented at The 16th Annual CUNY Conference on Human Sentence Processing.

Jäger, Lena, Zhong Chen, Qiang Li, Chien-Jer Charles Lin, and Shravan Vasishth (submitted). The subject-relative advantage in Chinese: Evidence for expectation-based processing.

Jurafsky, Daniel (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognition 20*, 137–194.

Just, Marcel, Patricia Carpenter, Timothy Keller, William Eddy, and Keith Thulborn (1996). Brain Activation Modulated by Sentence Comprehension. *Science 274*(5284), 114.

Kaplan, Tamar and John Whitman (1995). The Category of Relative Clauses in Japanese, with Reference to Korean. *Journal of East Asian Linguistics 4*(1), 29–58.

Kayne, Richard S. (1994). *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Keenan, Edward and Bernard Comrie (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry 8*(1), 63–99.

Keenan, Edward L. and Sarah Hawkins (1987). The psychological validity of the Accessibility Hierarchy. In E. L. Keenan (Ed.), *Universal Grammar: 15 Essays*, London, pp. 60–85. Croom Helm. The experimental work was carried out in England by Sarah Hawkins in 1974.

King, Jonathan and Marcel Just (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language 30*(5), 580–602.

King, Jonathan and Marta Kutas (1995). Who did what and when? Using word- and clause-level ERPS to monitor working memory usage in reading. *Journal of Cognitive Neuroscience 7*(3), 376–395.

Kurohashi, Sadao and Makoto Nagao (2003). Building a Japanese Parsed Corpus. In A. Abeillé and N. Ide (Eds.), *Treebanks*, Volume 20 of *Text, Speech and Language Technology*, pp. 249–260. Springer Netherlands.

Kwon, Nayoung, Yoonhyoung Lee, Peter C. Gordon, Robert Kluender, and Maria Polinsky (2010). Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of pre-nominal relative clauses in Korean. *Language 86*(3), 546–582.

Kwon, Nayoung, Maria Polinsky, and Robert Kluender (2006). Subject Preference in Korean. In D. Baumer, D. Montero, and M. Scanlon (Eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics (WCCFL-25)*, Somerville, MA, pp. 1–14. Cascadilla Proceedings Project.

Levelt, Willem (1974). *Formal grammars in linguistics and psycholinguistics*, Volume 192 of *Janua linguarum. Series minor*. The Hague: Mouton. Recently reprinted by John Benjamins isbn 978 90 272 3251 9.

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

Levy, Roger and Galen Andrew (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*.

Lewis, Richard L. and Shravan Vasishth (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science 29*, 1–45.

Lin, Chien-Jer Charles (2008). The processing foundation of head-final relative clauses. *Language and Linguistics 9.4*, 813–839.

Lin, Chien-Jer Charles (2014). Effect of thematic order on the comprehension of Chinese relative clauses. *Lingua 140*, 180–206.

Lin, Chien-Jer Charles (submitted). Subject prominence and processing dependencies in pronominal relative clauses: The comprehension of possessive relative clauses and adjunct relative clauses in Mandarin Chinese.

Lin, Chien-Jer Charles and Thomas Bever (2006). Subject preference in the processing of relative clauses in Chinese. In *Proceedings of the 25^{th} WCCFL*, pp. 254–260.

Lin, Chien-Jer Charles and Thomas Bever (2007). Processing doubly-embedded head-final relative clauses. In *Interdisciplinary Approaches to Relative Clauses*, Cambridge, UK.

Lin, Chien-Jer Charles and Thomas Bever (2011). Garden path and the comprehension of head-final relative clauses. In H. Yamashita, Y. Hirose, and J. L. Packard (Eds.), *Processing and Producing Head-final Structures,* Studies in Theoretical Psycholinguistics, pp. 277–297. Springer.

Lin, Yowyu Brian and Susan Garnsey (2011). Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin. In H. Yamashita, Y. Hirose, and

J. L. Packard (Eds.), *Processing and Producing Head-final Structures,* Studies in Theoretical Psycholinguistics, pp. 241–276. Springer.

MacDonald, Maryellen and Morten Christiansen (2002). Reassessing working memory: A reply to Just and Carpenter and Waters and Caplan. *Psychological Review 109*(1), 35–54.

MacWhinney, Brian (1977). Starting points. *Language 53*, 152–168.

MacWhinney, Brian (1982). Basic syntactic processes. In S. Kuczaj (Ed.), *Language Acquisition*, Volume 1 of *Syntax and Semantics*. Hillsdale, NJ: Lawrence Erlbaum.

Mak, Willem M., Wietske Vonk, and Herbert Schriefers (2002). The Influence of Animacy on Relative Clause Processing. *Journal of Memory and Language 47*(1), 50–68.

Mecklinger, Axel, Herbert Schriefers, Karsten Steinhauer, and Angela Friederici (1995). Processing relative clauses varying on syntactic and semantic dimensions: An analysis with event-related potentials. *Memory and Cognition 23*(4), 477–94.

Michaelis, Jens (2001). *On formal properties of Minimalist Grammars*. Ph. D. thesis, University of Potsdam, Potsdam, Germany.

Mitchell, Don, Fernando Cuetos, Martin Corley, and Marc Brysbaert (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research 24*, 469–488.

Miyamoto, Edson and Michiko Nakamura (2003). Subject/object asymmetries in the processing of relative clauses in Japanese. In *The 22nd West Coast Conference on Formal Linguistics (WCCFL-22)*, University of Californai, San Diego, pp. 342–355.

Miyamoto, Edson and Michiko Nakamura (2013). Unmet Expectations in the Comprehension of Relative Clauses in Japanese. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Nederhof, Mark-Jan and Giorgio Satta (2008). Computing partition functions of PCFGs. *Research on Language and Computation 6*, 139–162.

Ning, Chunyan (1993). *The Overt Syntax of Relativization and Topicalization in Chinese*. Ph. D. thesis, University of California, Irvine.

O'Grady, William (1997). *Syntactic Development*. The University of Chicago Press.

Packard, Jerome L., Zheng Ye, and Xiaolin Zhou (2011). Filler-gap processing in Mandarin relative clauses: Evidence from event-related potentials. In H. Yamashita, Y. Hirose, and J. Packard (Eds.), *Processing and Producing Head-final Structures,* Studies in Theoretical Psycholinguistics, Volume 38, pp. 219–240. Springer.

Qiao, Xiaomei, Liyao Shen, and Kenneth Forster (2012). Relative clause processing in Mandarin: Evidence from the maze task. *Language and Cognitive Processes 27*(4), 611–630.

Schachter, Paul (1973). Focus and relativization. *Language 49*, 19–46.

Schriefers, Herbert, Angela Friederici, and Katja Kühn (1995). The Processing of Locally Ambiguous Relative Clauses in German. *Journal of Memory and Language 34*, 499–520.

Sheldon, Amy (1974). On the role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior 13*, 272–281.

Smith, Raoul N. (1973). *Probabilistic Performance Models of Language*. Mouton.

Stabler, Edward (1997). Derivational minimalism. In C. Retoré (Ed.), *Logical Aspects of Computational Linguistics*. Springer-Verlag.

Stromswold, Karin, David Caplan, Nathaniel Alpert, and Scott Rauch (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language 52*, 452–473.

Traxler, Matthew, Robin Morris, and Rachel Seely (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language 47*, 69–90.

Ueno, Mieko and Susan Garnsey (2008). An ERP study of the processing of subject and object relative clauses in Japanese. *Language and Cognitive Processes 23(5)*, 646–688.

Vasishth, Shravan, Zhong Chen, Qiang Li, and Guilan Guo (2013). Processing Chinese Relative Clauses: Evidence for the Subject-Relative Advantage. *PLoS ONE 8*(10), e77006.

Vergnaud, Jean-Roger (1974). *French Relative Clauses*. Ph. D. thesis, Massachusetts Institute of Technology.

Wanner, E. and M. Maratsos (1978). An ATN approach in comprehension. In *Linguistic theory and psychological reality*, pp. 119–161. Cambridge:MA: MIT Press.

Whitman, John (2012). The prenominal relative clause problem. In U. Özge (Ed.), *Workshop on Formal Altaic Linguistics (WAFL) 8*. Cambridge, MA: MIT Working Papers in Linguistics (MITWPL).

Wilson, Kellogg and John B. Carroll (1954). Applications of entropy measures to problems of sequential structure. In C. E. Osgood and T. A. Sebeok (Eds.), *Psycholinguistics: a survey of theory and research*, pp. 103—110. Indiana University Press.

Wu, Fuyun (2009). *Factors Affecting Relative Clause Processing in Mandarin: Corpus and Behavioral Evidence*. Ph. D. thesis, University of Southern California.

Wu, Fuyun and Elsi Kaiser (submitted). Effects of early cues on the processing of Chinese relative clauses: Evidence for experience-based theories.

Wu, Fuyun, Elsi Kaiser, and Elaine Andersen (2012). Animacy effects in Chinese relative clause processing. *Language and Cognitive Processes 27*(10), 1489–1524.

Wu, Xiu-Zhi Zoe (2000). *Grammaticalization and the Development of Functional Categories in Mandarin*. Ph. D. thesis, University of Southern California.

Xue, Nianwen, Fei Xia, Fu-Dong Chiou, and Martha Palmer (2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering 11*(2), 207–238.

Yun, Jiwon, John Whitman, and John Hale (2010). Subject-object asymmetries in Korean sentence comprehension. In S. Ohlsson and R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.