Uncertainty reduction as a predictor of reading difficulty in Chinese relative clauses[*]

Zhong Chen *Cornell University*

Lena Jäger *University of Potsdam*

John Hale *Cornell University*

Abstract

This work applies the incremental complexity metric, Entropy Reduction (ER), to model the processing of prenominal Chinese relative clauses. ER formalizes the amount of information gained by an ideal parser as it strives to reduce structural uncertainty during sentence comprehension. In this work, per-word ER values reflect transformational syntactic proposals as well as distributional frequencies estimated from the Chinese Treebank. We show that these assumptions together are sufficient to derive the processing contrast between subject- and object-extracted relative clauses. This prediction is consistent with empirical results such as Lin & Bever 2006, among others. These ER values summarize changes in the probability distribution over syntactic alternatives, including expectations for future words. This approach enriches the structural frequency idea about relative clause processing with a more detailed explanation of ambiguity-resolution on a word-by-word basis.

Keywords: Sentence processing, Relative clauses, Computational modeling, Entropy Reduction

1. Introduction

A growing body of work in psycholinguistics suggests that it is possible to model incremental comprehension difficulties using information-theoretic notions (Hale, 2001, 2006; Levy, 2008). For instance, Entropy Reduction (ER, Hale, 2006) is a complexity metric that quantifies the cognitive effort expended on a word. The main idea is that words reduce uncertainty about the structure of the sentence. In this work, we apply the ER model to the processing of Chinese relative clauses (RCs), a construction that has long been studied in the psycholinguistic literature (Kaplan, 1974). ER predicts the extra reading difficulty in object relatives, consistent with experimental findings such as Lin & Bever 2006. This prediction in Chinese is also compatible with other RC modeling studies, e.g. English (Hale, 2006) and Korean (Yun, Whitman & Hale, 2010). It therefore suggests that human sentence processing is sensitive to both structural alternatives and their frequency distribution.

1.1 Disambiguation as Entropy Reduction

Entropy Reduction allows for the possibility of parallel parsing. The uncertainty that the ER deals with reflects the ambiguity between multiple parses, including

[*] The authors thank Tim Hunter and Jiwon Yun for their efforts in designing the modeling system. We are also indebted to Scott Chien-Hsiung Chiu, Charles Chien-Jer Lin, Shravan Vasishth and John Whitman for discussions. This paper is based upon work supported by the NSF CAREER Award No. 0741666 to John Hale. Correspondence can be addressed to zc77@cornell.edu .

expectations about as-yet-unheard words. As new words come in, given what one has already read, the probability of grammatical alternatives fluctuates. The idea of ER is that decreased uncertainty about the whole sentence, including probabilistic expectations, correlates with observed processing difficulty.[1] Such processing difficulty reflects the amount of information that a word supplies about the overall disambiguation task in which the reader is engaged.

The average uncertainty of specified alternatives can be quantified using the fundamental information-theoretic notion, entropy, as formulated below in definition (1). In a language-processing scenario, the random variable *X* in (1) might take values that are derivations on a probabilistic grammar *G*. We could further specialize *X* to reflect derivations proceeding from various categories, e.g., *NP*, *VP*, *S* etc. Since rewriting grammars always have a start symbol, e.g. *S*, the expression H_G(S) reflects the average uncertainty of guessing any derivation that *G* generates.

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

This entropy notation extends naturally to express conditioning events. If $w_1 w_{2\ldots}\, w_i$ is an initial substring of a sentence generated by *G*, the conditional entropy $H_G(\mathrm{S}|w_1 w_{2\ldots}\, w_i)$ will be the uncertainty about just those derivations that have $w_1 w_{2\ldots} w_i$ as a prefix.[2] By abbreviating $H_G(\mathrm{S}|w_1 w_{2\ldots}\, w_i)$ with H_i, the cognitive load *ER*(*i*) reflects the difference between conditional entropies before and after w_i, a particular word in a particular position in a sentence.

$$ER(i) = \begin{cases} H_{i-1} - H_i & \text{when this difference is positive} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Formula (2) defines the ER complexity metric. It says that cognitive work is predicted whenever uncertainty about the sentence's structure, as generated by the grammar *G*, goes down after reading in a new word.

Intuitively, disambiguation occurs when the uncertainty about the rest of the sentence decreases. In such a situation, readers' "beliefs" in various syntactic alternatives take on a more concentrated probability distribution (Jurafsky, 1996). The disambiguation work spent on this change is exactly the entropy reduction. By contrast when beliefs about syntactic alternatives become more disorganized, e.g. there exist many equiprobable syntactic expectations, then disambiguation work has not been done and the parser has gotten more confused. The background assumption of ER is that human sentence comprehension is making progress towards a peaked, disambiguated parser state and that the disambiguation efforts made during this process can be quantified by the reductions of structural uncertainty conveyed by words.

The ER proposal is not to be confused with another widely applied complexity metric, *Surprisal* (Hale, 2001; Levy, 2008), which is the conditional expectation of the log-ratio between forward probabilities of string prefixes before and after a word.

[1] The ER is a generalization of Narayanan and Jurafsky's (2002) idea of the "flipping the preferred interpretation". Here the flip only counts if the reader moves towards a less-confused state of mind.

[2] Conditional entropies can be calculated using standard techniques from computational linguistics such as chart parsing (Bar-Hillel, Perles & Shamir, 1964; Nederhof & Satta, 2008). See Chapter 13 of Jurafsky and Martin (2008) for more details.

Surprisal and ER have different properties, for instance entropy reductions are additive whereas surprisals are not.[3] ER has had rather better success in modeling sentence-medial ambiguities, such as those found in English object relatives. Surprisal has not led, as yet, to much insight into these effects (Levy, 2008: 1164).

As Section 3 goes on to show, the ER can provide a detailed account of ambiguity resolution in prenominal relativized structures like Chinese RCs. But before proceeding to the modeling itself, Section 2 first outlines the relevant empirical evidence.

2. Processing Relative Clauses

A prominent view (e.g. Fodor, 1978) supposes that a set of *universal processing principles* guide sentence processing in all human languages. Many researchers in the field have shared this idea and have instantiated it in a variety of ways. This section considers just one well-established processing pattern, the cross-linguistically attested preference for subject-extracted relative clauses and discusses a selection of universal processing principles that have been advanced as explanations for it.

2.1 Subject Relative Advantage

One robust processing pattern across languages is the subject and object asymmetry found in relative clause processing. In a relative clause, a noun phrase can be extracted/relativized from a variety of different "underlying" positions, for example, subject position or object position. The RC construction as a whole exhibits a Filler-Gap relationship. A large literature documents the finding that subject relatives (SRs) are easier to process than object relatives (ORs), a processing asymmetry known as the *Subject Advantage*. For example, in languages like English, the SR advantage has been observed in a variety of different measures, including: reading times (King & Just, 1991), eye-tracking (Traxler, Morris & Seely, 2002), ERP (King & Kutas, 1995), fMRI (Just, Carpenter, Keller, Eddy & Thulborn, 1996) and PET (Stromswold, Caplan, Alpert & Rauch, 1996). The subject advantage in relative clause processing has also been suggested in other languages, including those languages where relative clauses appear before the head noun (Lin, 2008).

A variety of more or less universal processing principles, as shown in Table 1, have been advanced as candidate explanations for the universal SR advantage pattern.[4] Among them, recent studies have appealed particularly to the working memory idea and to the structural frequency idea. The first explains the SR advantage in terms of a reduced memory load, compared to ORs whereas the second suggests that SRs are easier because they are more frequently used.

[3] Blachman (1968) clarifies the difference between surprisal and ER on a mathematical level. In his notation, "I" is the formal quantity that leads to ER and "J" is the quantity leading to surprisal.

[4] In some cases, the experimental results could not be explained by a single factor. For example, based on results of two eye movement experiments, Staub (2010) argues that both structural expectation-based accounts and memory retrieval-based accounts may co-determine the processing difficulty in English relative clauses. This proposal is also compatible with a computational modeling using reading data for naturally occurring relative clauses (Demberg & Keller, 2009).

Table 1. Processing principles proposed for relative clauses

Broad Categories		General Proposals
Word Order	Bever (1970); MacDonald & Christiansen (2002)	The sequence of words in SRs is closer to the canonical word order than that in ORs.
Accessibility Hierarchy	Keenan & Comrie (1977)	Universal markedness hierarchy of grammatical relations ranks the relativization from subject higher.
Working memory	Linear Distance: Wanner & Maratsos (1978); Gibson (2000); Lewis & Vasishth (2005)	ORs are harder than SRs because they impose a greater working memory burden.
	Structural Distance: O'Grady (1997); Hawkins (2004)	
Structural Frequency	Tuning Hypothesis: Mitchell, Cuetos, Corley & Brysbaert (1995); Jurafsky (1996)	SRs occur more frequently than ORs and therefore are more expected and easier to process.
	Surprisal: Hale (2001); Levy (2008)	ORs are more difficult because they require a low-probability rule.
	Entropy Reduction: Hale (2006)	ORs are harder because they force the comprehender through more confusing intermediate states.

2.2 Conflicting Results in Chinese

In the past decade, the SR advantage demonstrated in English and other languages has held up but not in every experiment. The processing-difficulty contrast between Chinese SRs and ORs, shown below in (3), is particularly interesting because previous studies have reported conflicting results.

(3) a. Subject Relatives
[e_i 邀請 富豪 的]$_{RC}$ 官員$_i$ 打了 記者
e_i invite tycoon de official$_i$ hit reporter
'The official who invited the tycoon hit the reporter.'

b. Object Relatives
[富豪 邀請 e_i 的]$_{RC}$ 官員$_i$ 打了 記者
tycoon invite e_i de official$_i$ hit reporter
'The official who the tycoon invited hit the reporter.'

In the above examples, the head noun "official" comes after the relative clause. As a result, the distance between the gap, indicated with a co-indexed empty category e_i, and the relativized head noun in SRs is longer than that in ORs in contrast to postnominal RCs in languages like English. This distance between filler and gap is particularly relevant to Working Memory theories based on linear distance (Table 1).

At first, these theories seemed to be confirmed by Chinese data. For example, Hsiao and Gibson (2003) reported that Chinese ORs are easier to comprehend than SRs. In particular, their experiment on single-embedded RCs like (3) showed that the two-word combination V+N in SRs is read slower than its counterpart in ORs, namely the N+V combination. They also observe slower reading times at the head noun in SRs but only in double-embedded RCs.

The irregular OR advantage could be explained by a working memory account: Dependency Locality Theory (DLT, Gibson, 1998, 2000). The DLT includes two processing cost metrics: the storage cost and the integration cost. The former explains the OR advantage in the sentence-initial two-word region whereas the latter accounts for the reading slow-down at SR's head noun.

The OR advantage reported by Hsiao and Gibson (2003) casts doubt on many other proposals for the processing pattern in RCs. For instance, the structural frequency idea explains the SR advantage in both the RC region and at the head noun, by appealing to the SR's higher construction frequency, compared to the OR. Comprehenders should have higher structural expectations on SRs because they occur more often than ORs, as suggested by corpus studies in a large number of languages.

Since the initial work by Hsiao and Gibson (2003), follow-up studies have yielded conflicting results. Chinese subject relatives are found to be either easier (C. Lin & Bever, 2006, 2011; C. Lin, 2008; Wu, 2009) or harder to process than object ones (Hsu & Chen, 2007; Y. Lin & Garnsey, 2011; Gibson & Wu, In Press). At present, the weight of the evidence seems to suggest that Chinese is not as exceptional as first suggested by Hsiao and Gibson (2003). However, a final determination awaits further investigation.

2.3 Processing the "Disambiguated" RCs

One of the problems with the experimental designs mentioned in the previous section is that there exist a number of temporary ambiguities in Chinese RCs. These ambiguities could affect the interpretation of observed processing patterns, no matter whether it is a subject advantage or an object one. Since Chinese RCs are prenominal, the head of an RC comes last. Syntactic alternatives could potentially compete with the RC reading. For example, when Hsiao and Gibson count the storage cost for Chinese SRs, they assume that the sentence is presented without context and therefore it is impossible for the sentence-initial verb to license a null subject such as an empty pronoun - *pro*. However, since subject *pro*-drop in Chinese is extremely frequent,[5] it is still possible that the processing of out-of-context experimental sentences may be influenced by other syntactic alternatives.

One natural way to solve this problem is to introduce a context preceding the critical sentence that promotes the expectation of an upcoming RC. Gibson and Wu (In Press) conducted such an experiment on subject-modifying single-embedded RCs.[6] They found a significant OR advantage at the head noun in contrast to the null result in

[5] A corpus search (Chen, Grove, Hale, In Press) in Chinese Treebank 7.0 (Xue et al. 2010) finds that there are 28913 simple sentences with a dropped subject whereas only 15996 simple sentences have an overt nominal subject (RC subjects are excluded). For the syntactic treatment of *pro*-drop in Chinese, see Huang, 1989 for details.

[6] Gibson and Wu (In Press) did not test object-modifying conditions perhaps because the main-clause illusion caused by the first three words (a N-V-N sequence) in object-modifying ORs cannot easily be eliminated even with a preceding context. (c.f. Lin & Bever 2011 for a discussion)

Hsiao & Gibson 2003. They interpreted this finding as evidence for the integration cost metric. However, they did not find the OR advantage at the RC region (the V+N in SR and N+V in OR) where a significant effect was reported in Hsiao & Gibson 2003.

Chen, Jäger, Li & Vasishth (Under Review), on the other hand, conducted a self-paced reading experiment on "disambiguated" RCs without the help from the preceding context. This experiment was designed to evaluate the opposing predictions of the frequency versus memory-based accounts.

(4) a. "Disambiguated" Subject Relatives
那個 昨晚 [e_i 揍了 服務生 一頓 的 $]_{RC}$ 顧客$_i$ 見過 老板 …
det-cl last night e_i hit-asp waiter one-cl de customer$_i$ see-asp boss …
'That customer who hit the waiter last night had seen the boss before…'

b. "Disambiguated" Object Relatives
那個 昨晚 [服務生 揍了 e_i 一頓 的 $]_{RC}$ 顧客$_i$ 見過 老板 …
det-cl last night waiter hit-asp e_i one-cl de customer$_i$ see-asp boss …
'That customer who the waiter hit last night had seen the boss before…'

Comparing example (4) with the regular RCs in (3), the head noun "customer" is now modified by a sentence-initial determiner-classifier combination. This sentence-initial sequence encourages readers to expect a noun phrase after processing the first segment of the sentence. However, the second segment of the sentence is a temporal phrase that could be attached either to a verb phrase or to a clause. This design leads the reader to foresee an upcoming relativized structure. In addition, the frequency/duration phrase before the relativizer "de" eliminates the possibility of treating the embedded noun phrase "waiter" in (2a) as a possessor of the head noun. It also increases the distance between the RC region and the head noun, which could be helpful in alleviating potential processing spillover from the RC region to the head noun (Vasishth & Lewis, 2003). As Chen et al. (Under Review) point out, the results indicate that SRs are read faster than ORs in the V+N or N+V region, at the head noun and the words afterwards. This pattern can be explained by Structural Frequency theories but not Working Memory theories.

Although the structural frequency idea is compatible with the SR advantage found in Chinese RCs and in other languages, as an explanation, it only goes so far. The intuition is that SRs occur more frequently and for this reason they are easier to comprehend. This intuitive idea can be instantiated in a variety of ways. For example, Chen et al. (In Press) uses surprisal to derive the Chinese SR advantage at the pre-head region but not at the head noun. In this paper, we employ a different frequency-related measure Entropy Reduction, which explains the reading slow-down at the head noun as well. The two both involve frequency but in different ways.

What ER measures is the reduction of uncertainty, in other words, the cognitive effort expended on disambiguation. How can it predict the correct processing pattern, e.g. reading times, for already "disambiguated" Chinese RCs? The answer is that sentences like (4) still have residual ambiguities. Although the presence of the determiner-classifier and the temporal phrase serve to reduce ambiguity greatly, there still exist other RC-like structures that could be viable alternatives to the globally correct analysis. The ER account is that reading times reflect, in part, the disambiguation of these alternatives. Because this disambiguation is defined over weighted syntactic alternatives, the overall proposal can be seen as a very particular type of structural frequency theory.[7]

3. Modeling

This section reports the procedure and results of modeling the processing of Chinese RCs. It suggests that by combining a formal grammar and structural frequency information, Entropy Reduction derives the observed pattern of comprehension difficulty on a word-by-word level.

3.1 Minimalist Grammar

Since structural alternatives are syntactic constructions, the first step of the modeling work is to prepare a grammar which covers the target sentences. We wrote a Minimalist Grammar (MG) in the style of Stabler (1997) for the Chinese RCs listed in example (4). It serves as the grammar *G* introduced in Section 1.1.

MGs are a transformational formalism that adopts ideas from the Minimalist Program (Chomsky, 1995). Stabler's formalization involves two generalized transformations: *Merge* and *Move*. Merge is a binary rule, analogous to ordinary context-free grammar rules or function application in categorial grammar. Move is a unary rule that is non-concatenative. Michaels (2001) and Harkema (2001) have shown that the languages generated by this system are mildly context-sensitive (Joshi, 1985). This means that while the derivation trees have a context-free, tree-like structure, and thus can be viewed as a weighted grammar, the derived languages exhibit the rich nonlocal dependencies, including crossing dependencies, that we find in natural language.

Switching from a context-free analysis of Chinese RCs (Chen et al., In Press) to one expressed within the MG formalism allows us to analyze the filler-gap dependency as feature-driven movement. The grammar fragment used in this work supposes that, in relativization, an argument NP moves to become the head of an RC. Table 2 lists a sample of six MG lexical items used in the Chinese grammar. For example, line 5 suggests that the relativizer *de* is provisionally treated as a complementizer analogous to English "that". The +f feature ensures that *de* selects a sentence to its left and a complementizer phrase (CP) is projected. The -k feature licenses a kind of movement that puts RCs on the left hand side of the NP they modify. In line 6, there is a nominal empty category *e* which takes a CP. It specifies that, after combining with its syntactic complement, this empty category raises a phrase out of the CP, a movement driven by the +wh feature. In a typical derivation this *wh*-movement will raise an argument noun phrase headed by the lexical entry in line 1.

The grammar uses abstract lexical items such as "Noun" or "Vt" so that entropy calculations based on it reflect only to structural uncertainty, as opposed to word-choice uncertainty. The grammar also differentiates RCs by extraction-site, i.e. $N_{\text{-SR}}$ and $N_{\text{-OR}}$. This is a case of *grandparent annotation* in the sense of Johnson (1998) to ensure that fine-grained probabilistic information can be captured in the grammar weighting stage. Using one category (e.g. $N_{\text{-RC}}$) instead of two categories $N_{\text{-SR}}$ and $N_{\text{-OR}}$ obscures this sort of distributional difference.

[7] The theory is obviously compatible with a Bayesian interpretation according to which structural probabilities are subjective beliefs rather than distributional frequencies.

Table 2. A sample of MG lexical entries used in the Chinese RC grammar

	Terminal Symbol	Syntactic Feature Sequence	Note
1.	Noun	N_{-Rel}, -case, -wh	relativizable noun with *wh*-feature
2.	*e*	$N_{-null-Rel}$, -case, -wh	null head has the same features as its overt counterpart
3.	Vt	=N, +case, V_{-SR}	transitive verb selecting a noun in an SR
4.	*e*	=>V_{-SR}, N_{-Rel}, v_{-SR}	verbal projection that selects a relativizable subject
5.	de	=T_{-SR}, +f, C_{-SR}, -k	relativizer *de* selects a sentence and projects a CP
6.	*e*	=C_{-SR}, +wh, N_{-SR}	empty head with *wh*-feature which relativizes a nominal

Our analysis follows Aoun and Li (2003) in the sense that in Chinese an NP rather than a DP is raised to the RC head position. The RC is then projected as an NP (c.f. Huang, Li & Li, 2009). In this way, the determiner and the classifier modifying the head noun can be outside of the RC. The MG focuses on the argument NP relativization. At this moment, it does not cover relativization involving resumptive pronouns or adjuncts. Therefore, for consistency with previous modeling work on prenominal Korean RCs (Yun et al., 2010), we employ the promotion analysis in Kayne's (1994) sense. Figure 1 shows an X-bar tree for the SR (4a) "Det Cl Time Vt Noun Freq de Noun Vt Noun", as generated by the Chinese grammar.

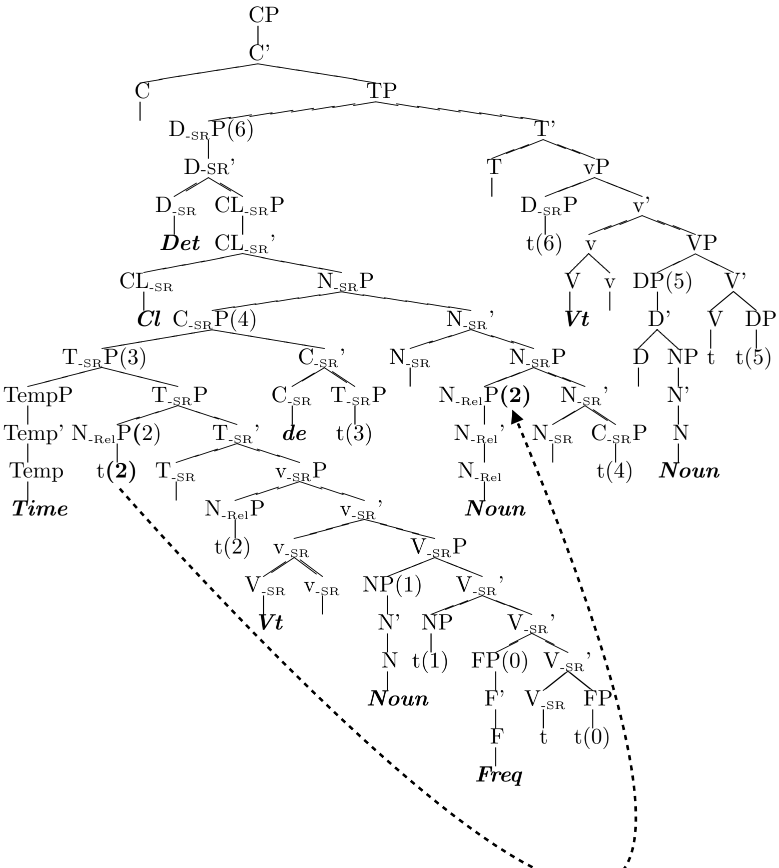


Figure 1. A derived SR generated by the MG of Chinese relatives

3.2 Grammar Weighting

The ER complexity metric derives processing difficulty, in part, from probabilities. This means we need to weight the prepared grammar with help from language resources.

The methodology of the grammar weighting is to estimate probabilistic context-free rules for MG derivations by parsing a "mini-Treebank." We obtain corpus counts for relevant structural types in the Chinese Treebank 7.0 (Xue et al., 2010) and treat these counts as attestation counts for particular constructions included in the mini-Treebank.[8] For example, Table 3 lists corpus counts of four RC constructions. It suggests that SRs are more frequent than ORs when they modify matrix subjects. In addition, ORs tend to allow more covert heads than SRs. By parsing the mini-Treebank, grammar rule weights are set by adding up the products of attestation counts and rule applications (Chi, 1999).

However, for structurally stricter sentence types such as the RC with a determiner-classifier and a frequency phrase in (4), we could not get enough corpus counts from the Chinese Treebank. We then estimate their counts proportionally based on their counterparts in a simpler version such as regular RCs in (3).

Table 3. A fragment of the "Mini-Treebank" that includes Chinese RC attestation rates

Strings	Construction Types	Corpus Counts
Vt Noun de Noun Vt Noun	Subject-modifying SR with Vt	366
Vt Noun de Vt Noun	Subject-modifying headless SR with Vt	123
Noun Vt de Noun Vt Noun	Subject-modifying OR with Vt	203
Noun Vt de Vt Noun	Subject-modifying headless OR with Vt	149

3.3 Prefix Parsing as Intersection Grammars

The weighted grammar allows us to calculate the probability of constructing a complete sentence. But, as Section 1.1 describes, the quantity that ER advances as a cognitive model is a conditional entropy. These values reflect uncertainty about every analysis and every grammatically possible sequence of words that can follow a given prefix string.

To compute these conditional entropies, we use chart parsing to recover probabilistic "intersection" grammars *G'* conditioned on each prefix of the sentences of interest (Nederhof & Satta, 2008). An intersection grammar derives all and only the sentences in the language of *G* that are consistent with the initial prefix. It implicitly defines comprehenders' expectations about how the sentence continues. Given the prefix string, the conditional entropy of the start symbol models a reader's degree of confusion about which construction he or she is in at that point in the sentence. Comparing the conditional entropies before and after adding a new word, any decrease quantifies disambiguation work that, ideally, could have been done at that word.

[8] Corpus inquiries about construction types were done by using the pattern-matching tool Tregex (Levy & Andrew, 2006). For an example of Tregex queries for Chinese RCs, see Table 4 in Chen et al. In Press.

Besides computing entropies, our system also samples syntactic alternatives from intersection grammars to get an intuitive picture of how uncertainties are reduced during parsing. These syntactic alternatives are discussed below in Section 3.4.

To summarize, the modeling procedure in the present work can be illustrated in a flowchart in Figure 2. By using a handwritten grammar, we focus on just a set of linguistically relevant alternatives. By using corpus counts, we attempt to define a realistic, frequency-sensitive notion of expectations in performance.

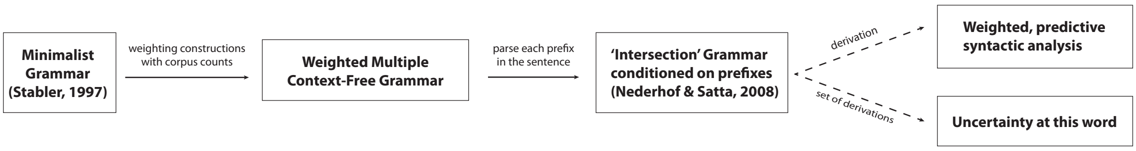

Minimalist Grammar (Stabler, 1997)
weighting constructions with corpus counts
Weighted Multiple Context-Free Grammar
parse each prefix in the sentence
'Intersection' Grammar conditioned on prefixes (Nederhof & Satta, 2008)
derivation
set of derivations
Weighted, predictive syntactic analysis
Uncertainty at this word

Figure 2. A modeling procedure illustration (adopted from Chen, Yun, Whitman & Hale, 2012)

3.4 Results

The ER predictions derive the subject-object asymmetry in Chinese RC examples (2). This derivation revises the frequency-based explanation suggested in Chen et al. Under Review. Figure 3 compares ER values per word in the sentence between an SR and an OR.

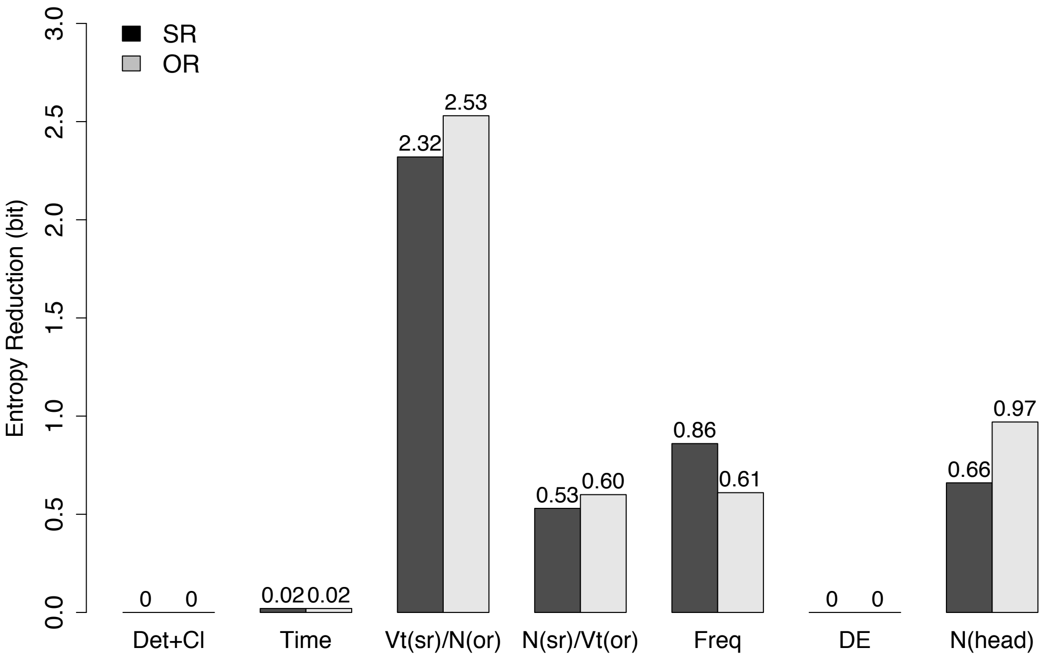
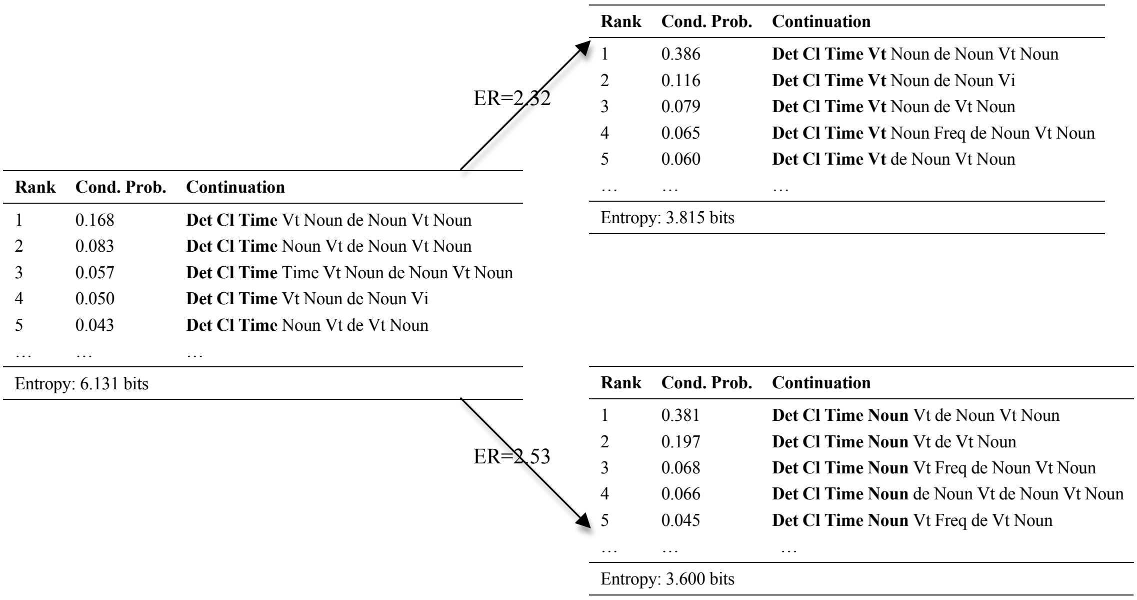


Figure 3. Entropy Reduction Predictions of Chinese RC Processing

Subject relatives have smaller ER values than object relatives at the third and fourth sets of words, namely the RC region (V+N in SR and N+V in OR). The SR advantage is more pronounced at the head noun position. In addition, the total ER value predicts an overall advantage on subject relatives.

The structural frequency idea is that SRs require less processing effort than ORs because they are more frequent. And indeed, the information-theoretical model discussed in this paper is consistent with this basic idea, i.e., the weighted grammar assigned probability 0.55 to SRs and 0.45 to ORs. However, to explain how pieces of relative frequency information actually guide the RC processing, one needs to carefully examine possible syntactic expectations at each prefix in the sentence. Our approach uses sampling to characterize what a reader would be expecting, given the grammar, at a particular point in the sentence. In most cases where a large ER difference is observed, this sampling procedure makes it straightforward to interpret entropy reductions in terms of re-orderings of the highest-value analyses.

Figure 4 explains how the subject advantage is predicted at the first set of words in the RC region, namely a transitive verb in SR and a noun in OR. Before reaching the RC, there is a common prefix for both the SR and the OR, a determiner-classifier combination followed by a temporal phrase "Det Cl Time" (bold in the figure). The conditional entropy calculated for this prefix is about 6.131 bits. This relatively high entropy value is determined by the fact that there is no dominating syntactic expectation with high conditional probability. In other words, at this point, in the sentence there is little reason to prefer one alternative over the other even though SRs in general are attested at a higher rate than ORs. Continuing the prefix with either a transitive verb (in SRs) or a noun (in ORs), the entropy reduction for these two transitions are different. It takes 2.53 bits ER to begin an OR whereas it costs only 2.32 bits to begin an SR. The conditional entropy calculated for the OR prefix "Det Cl Time Noun" is smaller than that of the SR prefix "Det Cl Time Vt", because in the OR two dominating syntactic alternatives emerge with probabilities 0.381 and 0.197 respectively. A more concentrated frequency distribution results in a less ambiguous state, and therefore a lower entropy value.



ER=2.32

Rank	Cond. Prob.	Continuation
1	0.386	**Det Cl Time Vt** Noun de Noun Vt Noun
2	0.116	**Det Cl Time Vt** Noun de Noun Vi
3	0.079	**Det Cl Time Vt** Noun de Vt Noun
4	0.065	**Det Cl Time Vt** Noun Freq de Noun Vt Noun
5	0.060	**Det Cl Time Vt** de Noun Vt Noun
…	…	…
Entropy: 3.815 bits		

Rank	Cond. Prob.	Continuation
1	0.168	**Det Cl Time** Vt Noun de Noun Vt Noun
2	0.083	**Det Cl Time** Noun Vt de Noun Vt Noun
3	0.057	**Det Cl Time** Time Vt Noun de Noun Vt Noun
4	0.050	**Det Cl Time** Vt Noun de Noun Vi
5	0.043	**Det Cl Time** Noun Vt de Vt Noun
…	…	…
Entropy: 6.131 bits		

ER=2.53

Rank	Cond. Prob.	Continuation
1	0.381	**Det Cl Time Noun** Vt de Noun Vt Noun
2	0.197	**Det Cl Time Noun** Vt de Vt Noun
3	0.068	**Det Cl Time Noun** Vt Freq de Noun Vt Noun
4	0.066	**Det Cl Time Noun** de Noun Vt de Noun Vt Noun
5	0.045	**Det Cl Time Noun** Vt Freq de Vt Noun
…	…	…
Entropy: 3.600 bits		

Figure 4. ER at the first set of words in Chinese RCs

Earlier modeling work on Chinese RCs (Chen et al, In Press) did not derive the SR advantage at the RC final head noun (Lin & Bever, 2006; Chen et al., Under Review). The current work with Entropy Reduction does this. It suggests that the frequency of "un-chosen" null-heads plays an important role in explaining the ease of processing an

SR's head noun. By examining pre-head syntactic expectations, as shown in Figure 5, we find that in ORs there is at least a 32% chance that the prefix will continue as a headless RC. On the contrary, it is less likely that an SR prefix will have a covert head. This contrast, consistent with corpus counts in Table 3, suggests that integrating an overt head into the prenominal structure is easier in SRs than in ORs because less uncertainty about the overall structure is eliminated.

Rank	Cond. Prob.	Continuation
1	0.552	**Det Cl Time Vt Noun Freq de** Noun Vt Noun
2	0.166	**Det Cl Time Vt Noun Freq de** Noun Vi
3	0.112	**Det Cl Time Vt Noun Freq de** Vt Noun
4	0.058	**Det Cl Time Vt Noun Freq de** Noun Vt Noun de Noun
5	0.034	**Det Cl Time Vt Noun Freq de** Vi
…	…	…
Entropy: 2.406 bits		

ER=0.66

Rank	Cond. Prob.	Continuation
1	0.664	**Det Cl Time Vt Noun Freq de Noun** Vt Noun
2	0.200	**Det Cl Time Vt Noun Freq de Noun** Vi
3	0.069	**Det Cl Time Vt Noun Freq de Noun** Vt Noun de Noun
4	0.008	**Det Cl Time Vt Noun Freq de Noun** Vt Det Cl Vt Noun de Noun
5	0.008	**Det Cl Time Vt Noun Freq de Noun** Vt Vt Noun de Noun
…	…	…
Entropy: 1.750 bits		

Rank	Cond. Prob.	Continuation
1	0.490	**Det Cl Time Noun Vt Freq de** Noun Vt Noun
2	0.320	**Det Cl Time Noun Vt Freq de** Vt Noun
3	0.051	**Det Cl Time Noun Vt Freq de** Noun Vt Noun de Noun
4	0.033	**Det Cl Time Noun Vt Freq de** Vt Noun de Noun
5	0.015	**Det Cl Time Noun Vt Freq de** Noun Vi
…	…	…
Entropy: 2.390 bits		

ER=0.97

Rank	Cond. Prob.	Continuation
1	0.810	**Det Cl Time Noun Vt Freq de Noun** Vt Noun
2	0.085	**Det Cl Time Noun Vt Freq de Noun** Vt Noun de Noun
3	0.025	**Det Cl Time Noun Vt Freq de Noun** Vi
4	0.010	**Det Cl Time Noun Vt Freq de Noun** Vt Det Cl Vt Noun de Noun
5	0.010	**Det Cl Time Noun Vt Freq de Noun** Vt Vt Noun de Noun
…	…	…
Entropy: 1.422 bits		

Figure 5. ER at the head noun in Chinese RCs

4. Conclusion

In this work, Entropy Reduction, in conjunction with a formal grammar weighted by corpus counts, models the subject advantage in Chinese relative clause processing. This result is consistent with the intuitive structural frequency idea, namely that a frequent structure is easier to comprehend. However, it takes this idea further by highlighting the particular disambiguation decisions that contribute to predicted difficulty. These predictions are consistent, at a region-by-region level, with data collected by Chen et al. (Under Review). We suspect that an even finer-grained explanation might follow from a model that incorporates additional information. One such factor to be considered in future work is animacy, as suggested by Wu, Kaiser & Andersen (In Press).

References

Aoun, Joseph, and Yen-Huei Audrey Li. 2003. *Essays on the representational and derivational nature of grammar: the diversity of wh-constructions*. Cambridge, MA: MIT Press.

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

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

Bever, Thomas G. 1970. The cognitive basis for linguistic structures. Hayes, J. R. ed., *Cognition and the Development of Language*, John Wiley, 279–360.

Chen, Zhong, Kyle Grove, and John Hale. In Press. Structural expectations in Chinese relative clause comprehension. In *Proceedings of the 29th West Coast Conference on Formal Linguistics*, Somerville, MA: Cascadilla Proceedings Project.

Chen, Zhong, Lena Jäger, Qiang Li, and Shravan Vasishth. Under Review. The subject-relative advantage in Chinese relatives: Evidence for expectation-based processing.

Chen, Zhong, Jiwon Yun, John Whitman, and John Hale. 2012. Uncertainty and Prediction in Relativized Structures across East Asian Languages. Poster presented at *the 25th Annual CUNY Conference on Human Sentence Processing*, New York, NY, March 14-16.

Chi, Zhiyi. 1999. Statistical properties of probabilistic context-free grammars. *Computational Linguistics* 25(1), 131-160

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

Demberg, Vera, and Frank Keller. 2009. A Computational Model of Prediction in Human Parsing: Unifying Locality and Surprisal Effects. In *Proceedings of the 29th meeting of the Cognitive Science Society*, Amsterdam, Netherlands.

Fodor, Janet. 1978. Parsing Strategies and Constraints on Transformations. *Linguistic Inquiry*, 9(3), 427-473

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

Gibson, Edward. 2000. Dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, brain: Papers from the First Mind Articulation Project Symposium*, ed. by Alec Marantz, Yasushi Miyashita and Wayne O'Neil, MIT Press, Cambridge, MA.

Gibson, Edward, and Hsiao-Hung Iris Wu. In Press. Processing Chinese relative clauses in context. *Language and Cognitive Processes*.

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

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

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd North American Association of Computational Linguistics* 159-166, Pittsburg, PA.

Hale, John. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30.4: 643-672.

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

Hsu, Chun-Chieh, and Jenn-Yeu Chen. 2007. A new look at the subject-object asymmetry: The effects of linear distance and structural distance on the processing of head-final relative clauses in Chinese. *Interdisciplinary Approaches to Relative Clauses*, Cambridge, UK.

Huang, C.-T. James. 1989. Pro-drop in Chinese: a generalized control approach. In *The null subject parameter*, ed. Osvaldo Jaeggli and Ken Safir, 185-214. Dordrecht: D. Reidel.

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

Johnson, Mark. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*, 24: 613–632.

Joshi, Aravind K. 1985. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In D. Dowty, L. Karttunen, and A. M. Zwicky eds., *Natural language parsing: Psychological, computational and theoretical perspectives* (pp. 206–250). New York: Cambridge University Press.

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

Jurafsky, Daniel, and James H. Martin, 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Second Edition, McGraw Hill

Just, Marcel A., Patricia A. Carpenter, Timothy A. Keller, William F. Eddy, and Keith R. Thulborn. 1996. Brain Activation Modulated by Sentence Comprehension. *Science*. 274: 5284, p. 114.

Kaplan, Ronald. M. 1974. *Transient processing load in relative clauses*. Unpublished doctoral dissertation, Harvard University.

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, 63-99.

King, Jonathan, and Marcel A. Just. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30, 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.

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

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)*, Genoa.

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

Lin, Chien-Jer Charles, and Thomas G. Bever. 2006. Subject preference in the processing of relative clauses in Chinese. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, ed. by Donald Baumer, David Montero and Michael Scanlon, Cascadilla Proceedings Project, Sommerville, MA, 254-260.

Lin, Chien-Jer Charles, and Thomas G. Bever. 2011. Garden path and the comprehension of head-final relative clauses. In Yamashita, Hiroko, Yuki Hirose and Jerome L. Packard eds., *Processing and Producing Head-final Structures*, Studies in Theoretical Psycholinguistics, Springer, 277–297.

Lin, Yowyu Brian, and Susan M. Garnsey. 2011. Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin. In Yamashita, Hiroko, Yuki Hirose & Jerome L. Packard eds., *Processing and Producing Head-final Structures*, Studies in Theoretical Psycholinguistics, Springer, 241–276.

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

Michaelis, Jens. 2001. *On formal properties of minimalist grammars*. Ph.D. thesis, University of Potsdam, Potsdam, Germany.

Mitchell, Don C., Fernando Cuetos, Martin M. B. 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.

Narayanan, Srini, and Daniel Jurafsky. 2002. A Bayesian Model Predicts Human Parse Preference and Reading Time in Sentence Processing. In T. G. Dietterich, S. Becker and Z. Ghahramani, eds., *Advances in Neural Information Processing Systems* 14. Cambridge, MA: MIT Press. 59-65.

Nederhof, Mark-Jan, and Satta Giorgio. 2008. Computing Partition Functions of PCFGs. *Research on Language and Computation*, 6.2: 139–162.

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

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

Stabler, Edward. 1997. Derivational minimalism. *Logical aspects of computational linguistics*, ed. by Christian Retoré. 68-95. NY: Springer-Verlag. (Lecture Notes in Computer Science 1328)

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

Stromswold, Karin, David Caplan, Nathaniel Alpert, and Scott Rauch. 1996. Localization of Syntactic Comprehension by Positron Emission Tomography. *Brain and Language*, 52:3, 452–473.

Traxler, Matthew J., Robin K. Morris, and Rachel E. Seely (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47.1, 69-90.

Wanner, Eric, and Michael Maratsos. 1978. An ATN approach to comprehension. In Morris Halle, Joan Bresnan, and George A. Miller, eds, *Linguistic Theory and Psychological Reality*, 119–161. Cambridge, MA: MIT Press.

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

Wu, Fuyun, Elsi Kaiser, and Elaine Andersen. In Press. Animacy effects in Chinese relative clause processing. *Language and Cognitive Processes*.

Xue, Nianwen, Zixin Jiang, Xiuhong Zhong, Martha Palmer, Fei Xia, Fu-Dong Chiou, and Meiyu Chang. 2010. *Chinese Treebank 7.0*. Linguistic Data Consortium.

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*.