Abstract

The Entropy Reduction Hypothesis (Hale, 2006) derives the subject-object asymmetry in Korean relative clauses. This asymmetry has been observed by Kwon, Polinsky, and Kluender (2006), among others. Agreement between the Entropy Reduction predictions and the available empirical data suggests that the heightened comprehension difficulty attested in object-extracted relatives is due to distinctive incremental parser states associated with comparatively greater temporary ambiguity.

Keywords: sentence comprehension, relative clauses, Korean, probabilistic grammar, Entropy Reduction, syntax

Introduction

Relative clauses (RCs) have long been objects of fascination for cognitive scientists interested in language comprehension (Kaplan, 1974). In the well-known “subject-extracted” (SRC) and “object-extracted” (ORC) cases, a large literature exists. In languages such as English and French, a processing advantage for SRCs has been confirmed in a wide variety of measures including phoneme-monitoring (Frauenfelder, Segui, & Mehler, 1980), eye-fixations (Holmes & O'Regan, 1981), reading times (King & Just, 1991), PET (Stromswold, Caplan, Alpert, & Rauch, 1996) and fMRI (Just, Carpenter, Keller, Eddy, & Thulborn, 1996). It has been suggested that the SRC advantage may be a processing universal (Lin, 2008). If ORCs are harder than SRCs in all languages, then what is it about human sentence comprehension that makes this so? The Korean language is a key test for any universal processing theory because it is syntactically different from English and French. These differences include verbal-final clauses and prenominal RCs.

In this paper, we offer an account of the SRC/ORC asymmetry in terms of the information-processing difficulty of incremental parsing in general. This proposal relates the hardness of parsing to syntactic facts about Korean. A language independent complexity metric known as Entropy Reduction (Wilson & Carroll, 1954; Hale, 2003, 2006) correctly derives the SRC advantage when applied with a Korean grammar. This demonstration supports the claim that human comprehension difficulty reflects the kind of information-processing work that Entropy Reduction quantifies.  

Theories of the Subject-Object Asymmetry

As an empirical phenomenon, the SRC/ORC processing asymmetry is well-established. However, its implications for the architecture or mechanisms of human language comprehension remain controversial. Three broad classes of theory have been advanced. LINEAR DISTANCE theories, illustrated in Figure 1, point to a greater number of intervening elements between the relativized position and the headnoun to which it is meaningfully related. The boxed e notation stands for an “empty” element. Particular theories of LINEAR DISTANCE offer alternative ways of measuring the separation between this omitted position and the headnoun (Wanner & Maratsos, 1978; Gibson, 2000; Lewis & Vasishth, 2005). These theories all provide an adequate account of the English pattern, and in some cases relate this prediction to plausible mechanisms of human sentence comprehension. They are thwarted, however by data that confirm an SRC-over-ORC processing advantage in Korean (O’Grady, Lee, & Choo, 2003; Kwon et al., 2006; Lee, 2007). Figure 1(b) shows how theories of this type derive the wrong prediction for Korean.

The second broad class includes STRUCTURAL DISTANCE theories. The simplest theory of this kind maintains that ORCs are harder because the relativized element is more deeply embedded when it is an object. If ORCs are formed by a movement rule, then this movement would “cross” both a VP node and an S node to arrive at its surface position (O’Grady, 1997, 179). Hawkins (2004, 175) singles-out “a connected path that must be accessed for gap identification and processing.” Hawkins’ path is shown using dotted branches in Figure 2. This path is shorter for SRCs in both Korean and English. This general account is thus adequate but not very precise. It leaves open, for instance, the question of where exactly greater difficulty should start to accrue during incremental processing.

The third broad class contains the INFORMATION-THEORETICAL approaches. The Entropy Reduction Hypothesis (ERH) fits into this class. It holds that a person’s difficulty at a word reflects the amount by which that word helped him or her to ascertain which construction the speaker intends. The ERH uses the concept of entropy to quantify the average uncertainty about derivations consistent with an observed initial string. This entropy is high when there are many equiprobable continuations and low when there are just a few continuations or the probability distribution on them is sharply concentrated. This quantity standsin for the degree of confusion in the comprehender’s mind. When it is reduced in the transition from one word to the next, the comprehender...
has accomplished disambiguation work. The ERH interprets this theoretical work as a word-by-word metric of incremental comprehension difficulty.

Hale (2006) derives Entropy Reduction predictions for English relative clauses. Asymmetries between them suggest that relativized non-subjects are harder to comprehend because of greater temporary ambiguity at the embedded verb. While it is well-known that Korean exhibits considerable temporary ambiguity in the middle of sentences, precise values have not been compared across constructions. Figure 3 illustrates this ambiguity with a prefix string that could signal at least four different clause-types. The ERH offers the possibility of accounting for the SRC/ORC asymmetry in terms of contrasting levels of such ambiguity.

**Procedure**

We calculate Entropy Reductions at every inter-word point in Korean SRC and ORC sentences using a procedure that mirrors Hale (2006). One of us (JY) prepared a Korean grammar that covers the sentences listed in the Appendix. This grammar is written in Stabler’s Minimalist Grammars (MG) formalism (Stabler, 1997). This transformational formalism adopts certain themes of Chomsky’s Minimalist Program (1995) and has been shown to be mildly context-sensitive in the sense of Joshi (1985) by Michaelis (2001). We consider subject-extraction and object-extraction in each of the four clause-types shown in Figure 3. Our analysis supposes that the headnoun moves in relativization. We use the MG move rule to implement this analysis. Figure 4 shows a structural description generated by this grammar. This grammar analyzes postnominal case markers as separate words and verb suffixes as part of verbs. Here, a coindexed trace, t(3) indicates movement of the headnoun kica ‘reporter’ from its base position in a specifier of little v to a position outside the RC. Weighting each construction type listed in the Appendix by its attestation count in a Korean Treebank (Han et al., 2006), we estimate a probabilistic context-free grammar (PCFG) of MG derivations. By chart parsing, we recover a new PCFG for each prefix of the sentences of interest. This chart-PCFG is an alternative presentation of the AND-OR graph encoded by the chart (Lang, 1991). It represents all possible analyses that are consistent with the given prefix. We calculate the entropy of the start symbol of this chart-PCFG to arrive at the conditional entropy of the prefix string. This value is a cognitive model of an incremental comprehender’s degree of confusion about which construction he or she is in. When it goes down, disambiguation work has occurred.

**Results**

Table 1 summarizes the ERH predictions: SRCs are easier to comprehend than ORCs. This prediction also follows in noun complement clauses. However, empty elements in subject position are not always easier. In simple matrix clauses and adjunct clauses, no difference is predicted.

<table>
<thead>
<tr>
<th>Clause type</th>
<th>SBJ Extraction</th>
<th>OBJ Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix Clause</td>
<td>19.6</td>
<td>19.6</td>
</tr>
<tr>
<td>Adjunct Clause</td>
<td>34.66</td>
<td>34.66</td>
</tr>
<tr>
<td>Complement Clause</td>
<td>32.1</td>
<td>42.98</td>
</tr>
<tr>
<td>Relative Clause</td>
<td>27.13</td>
<td>35.65</td>
</tr>
</tbody>
</table>

Table 1: Average Entropy Reduction in bits-per-word
Word-by-word Entropy Reduction graphs, shown in Figure 6, illustrate how predicted difficulty peaks coincide with the positions that disambiguate clause-type and the role of omitted elements. This is indicated with double-circles in Figure 5. The subject-object asymmetry in RCs is predicted to show up on the headnoun at the position marked N in Figure 6(d). This prediction matches the findings of Kwon et al. (2006), who observe a reading time asymmetry at this point.

**Discussion**

The Entropy Reduction account of the subject advantage in relative clauses and complement clauses is rooted in contrasting levels of uncertainty about syntactic structure. The crucial position, immediately after the adnominal form of the verb, is marked ỹ in Figure 7. In the ORC case, the conditional entropy at this point is 32.28 bits, while in the SRC case, the corresponding value is only 23.76 bits. The conditional entropy values at ỹ are exactly the same — 17.43 bits in both cases. Thus, the ERH models the greater difficulty in the object cases with greater conditional entropy at point ỹ.

The disparity between these conditional entropies reflects contrasting numbers of alternative continuations. These continuations correspond to different roles the prefix string might play at the matrix level. Figure 8 shows that the ORC prefix \textsc{N NOM V-ADN} could be in fact the beginning of a reading on which the nominative-marked noun is a complete matrix-level subject on its own, where both the subject and the object of the embedded clause are omitted. These properties allows the ORC prefix to have the multiple parses shown in (1-3) below. The disparity derives, ultimately, from syntactic properties of Korean. As we have seen, it is an SOV language with prenominal RCs; crucially, arguments may be freely omitted when they are recoverable in-context. Such additional structures are not acceptable as a continuation of the SRC prefix \textsc{N ACC V-ADN}, which cannot be split by additional empty categories.

\begin{equation}
\text{kica ka } [\text{SRC } \text{kongkyekhan}] \text{ uywon ul manassta.}
\end{equation}
Figure 5: Continuations signal clause-types

Figure 6: Word-by-word comprehension difficulty predictions derived by the INFORMATION-THEORETICAL Entropy Reduction Hypothesis. Horizontal axes labels name word classes. SBJ abbreviates “subject-extracted”, OBJ “object-extracted”. Clause-types (a)–(d) are as in Figure 3.
they do include the prediction of a subject-object asymmetry in complement clauses with omitted arguments. The effect should appear on the word sasil ‘fact’. This prediction would not follow on a STRUCTURAL DISTANCE account, since no movement relation exists between the empty element pro and sasil in that construction. If a subject-object asymmetry were to be experimentally observed at that point, this would leave the ERH as the only theory able to explain the English as well as the Korean results. We hope that our work encourages empirical investigation of this case.

Acknowledgment
The authors would like to express their appreciation to Nayoung Kwon for her comments on the manuscript. Thanks are due as well to Na-Rae Han for information about the Korean Treebank. This research was supported by a Small Grant from the Cornell University Institute for the Social Sciences.

Appendix: Examples
The Minimalist Grammar used to derive the comprehension-difficulty predictions graphed in Figure 6 covers all of the examples listed below. The combinatorics of the promotion analysis imply the existence of other grammatical strings such as the examples (1)–(3) in discussion.

a. matrix clause with a pro-subject

\[ uywon \text{ kongkyekhaysesta.} \]

senator ACC attack-DECL

‘Someone attacked the senator.’

b. adjunct clause with a pro-subject

\[ uywon \text{ kongkyekhayse kica ka} \]

senator ACC attack-ADV reporter NOM

\[ yamyenghaycyessta. \]

become-famous-DECL

‘Because someone/he attacked the senator, the reporter became famous.’

c. complement clause with a pro-subject

\[ uywon \text{ kongkyekhan sasil i palkhyecyessta.} \]

senator ACC attack-ADV fact NOM is-revealed-DECL

‘The fact that someone attacked the senator was revealed.’

d. subject relative clauses

Related work
These results offer a new perspective on the work of Ishizuka, Nakatani, and Gibson (2006). Using Japanese RCs, which are structurally similar to Korean, these authors show that the penalty for ORC processing can be mitigated or even eliminated if certain readings are pragmatically suppressed by prior discourse. The ERH suggests that disambiguating those readings is exactly the source of the ORC penalty. It quantifies the difficulty of coping with all the available alternatives.

Our results also suggest a lack of subject-object asymmetry in adjunct clauses. We would like to emphasize that this does not entail a contradiction with the experimental results of Kwon et al. (2006). The design of this experiment leverages the fact that a matrix clause noun is a felicitous controller of pro when it appears in an embedded clause. Indeed, these authors suggest that “the identification of the gap in an adjunct clause does not involve any syntactic operations.” It is thus appropriate that our syntax-only approach predicts no distinction between missing subjects and objects in this clause type. The ERH might naturally be combined with a pragmatic component to yield a broader theory. We leave this extension to future work.

Conclusion
The ERH, in conjunction with an appropriate formal grammar, can account for the subject advantage in Korean RCs. Its predictions cannot be summarized by simply saying that missing objects are always harder; for instance both types of main clauses are predicted to be equally easy. However
Hale, J. (2003). The information conveyed by words in sen-
g. object relative clauses
kica ka kongkyekhan uywon i
reporter NOM attack-ADV senator NOM
yummyenghaycyessta.
become-famous-DECL.

‘The reporter whom the reporter attacked became famous.’

h. object relative clauses
kica ka kongkyekhan uywon i
reporter NOM attack-ADV senator NOM
yummyenghaycyessta.
become-famous-DECL.

‘The senator whom the reporter attacked became famous.’

References
Verbal Behavior, 19(3), 328-337.
Hale, J. (under review). What a rational parser would do.
Michaelis, J. (2001). Derivational minimalism is mildly context-sensitive. In M. Moortgat (Ed.), Logical aspects of computational linguistics (pp. 179-198). Springer. (Selected papers from LACL98)