Subject-Object Asymmetries in Korean Sentence Comprehension

Jiwon Yun (jy249@cornell.edu)
John Whitman (jbw2@cornell.edu)
John Hale (jthale@cornell.edu)
Department of Linguistics, 203 Morrill Hall, Cornell University
Ithaca, NY 14850 USA

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 verb-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. [1]

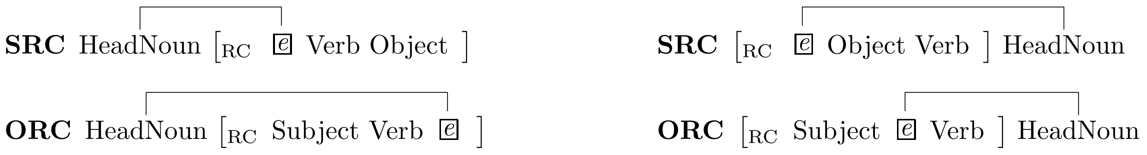
Theories of the Subject-Object Asymmetry

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

[1] A longer companion paper, Hale (under review), develops an automaton model of the sentence comprehension process. It presents a generalized left-corner parser that operates in accordance with the Entropy Reduction Hypothesis when its decisions about how to resolve nondeterminism are guided by experience.


SRC HeadNoun [RC e Verb Object]
ORC HeadNoun [RC Subject Verb e]
SRC [RC e Object Verb] HeadNoun
ORC [RC Subject e Verb] HeadNoun

(a) correct SRC<ORC prediction for English

(b) unsupprted SRC>ORC prediction for Korean

Figure 1: Predictions of LINEAR DISTANCE

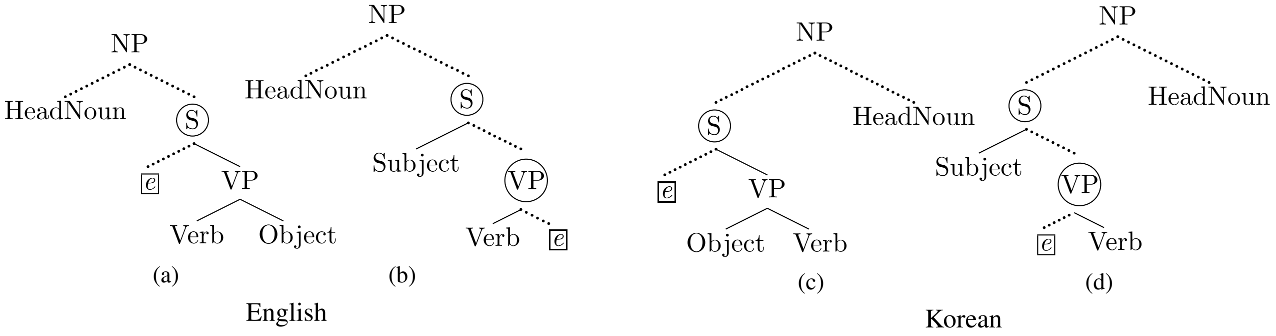


Figure 2: Predictions of STRUCTURAL DISTANCE. In ORCs, ((b),(d)) the pathway between e and HeadNoun crosses two circled nodes whereas in SRCs it crosses just one ((a),(c)). This asymmetry makes the right prediction in both languages.

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 levels 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.

Clause type	SBJ Extraction	OBJ Extraction
Matrix Clause	19.6	19.6
Adjunct Clause	34.66	34.66
Complement Clause	32.1	42.98
Relative Clause	27.13	35.65

Table 1: Average Entropy Reduction in bits-per-word

matrix clause ⓔ ***uywon ul*** *kongkyekhayssta* pro senator ACC attack-DECL '(someone) attacked the senator.'	complement clause ⓔ ***uywon ul*** *kongkyekhan sasil* pro senator ACC attack-ADN fact 'the fact that (someone) attacked the senator'
adjunct clause ⓔ ***uywon ul*** *kongkyekhayese* pro senator ACC attack-ADV 'because (someone) attacked the senator,'	relative clause ⓔ ***uywon ul*** *kongkyekhan kica* gap senator ACC attack-ADN reporter 'the reporter who attacked the senator'

Figure 3: The same initial morphemes signal at least four different clause types[2]

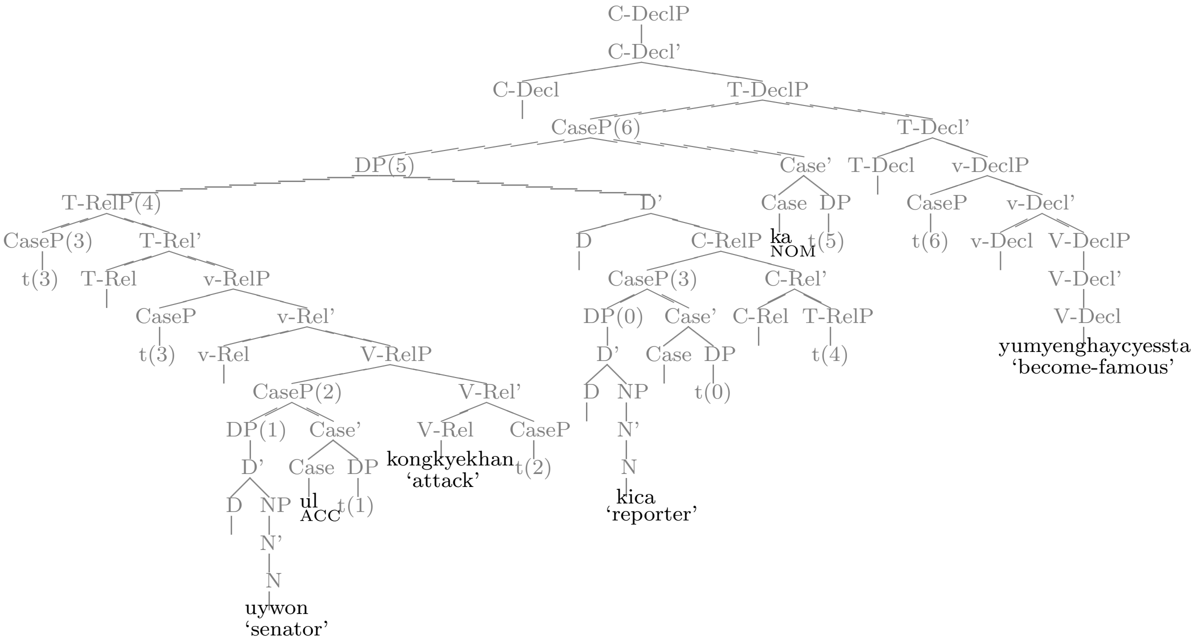


Figure 4: Structural description of SRC example (d) from the Appendix

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 ❸ 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 ④ 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 ❸.

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 **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 **N ACC V-ADN**, which cannot be split by additional empty categories.

(1) *kica ka* $[_{\text{SRC}}$ ⓔ ⓔ *kongkyekhan* $]$ *uywon ul*
reporter NOM gap pro attack-ADN senator ACC
manassta.
meet-DECL

[2] Our notational conventions include NOM for nominative case, ACC for accusative, ADV for adverbial, ADN for adnominal and DECL for declarative.

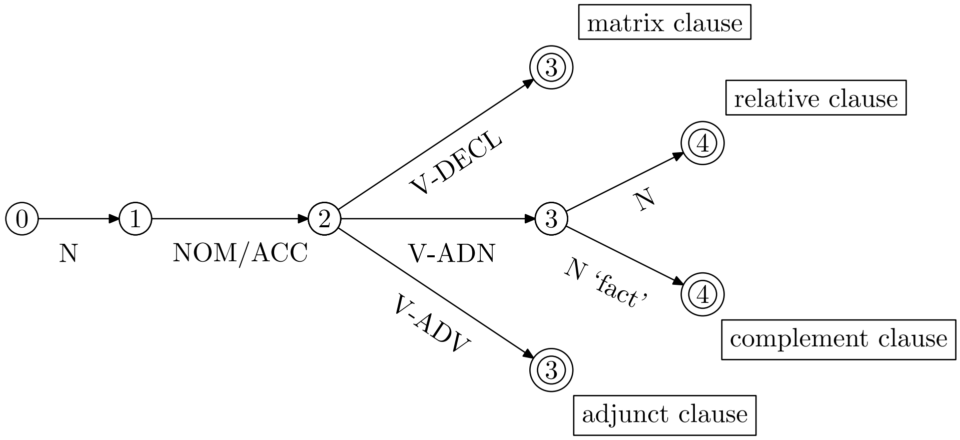
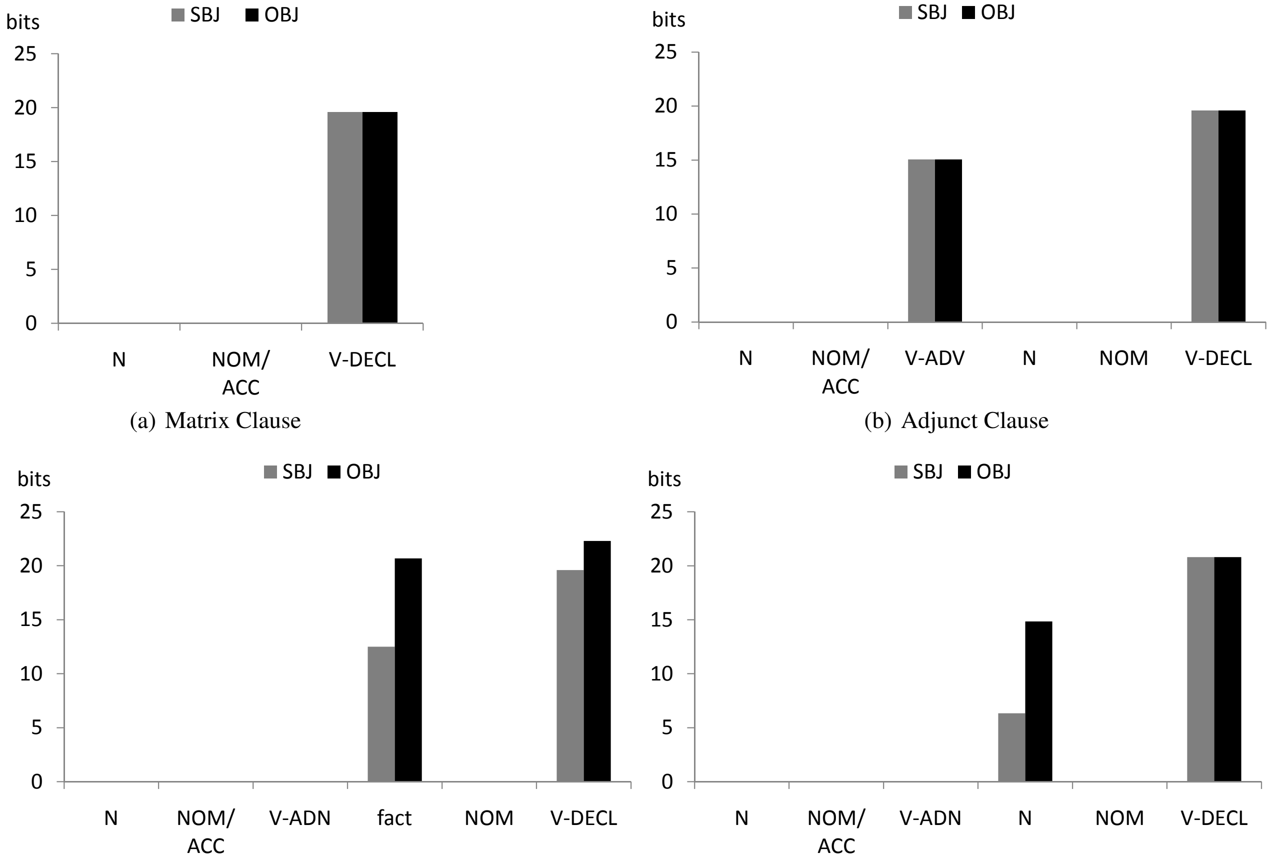


Figure 5: Continuations signal clause-types



(a) Matrix Clause

(b) Adjunct Clause

(c) Complement Clause

(d) Relative Clause

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.

SRC *kica* ① *lul* ② *kongkyekhan* ❸ *uywon* ④
reporter ACC attack-ADN senator
'the senator who attacked the reporter'

ORC *kica* ① *ka* ② *kongkyekhan* ❸ *uywon* ④
reporter NOM attack-ADN senator
'the senator who the reporter attacked'

Figure 7: SRC and ORC. The black circle indicates where the difference of structural uncertainty is observed.

SRC a. [$\boxed{e}$ *kica lul kongkyekhan*]
[SBJ OBJ V-ADN]

ORC a. [*kica ka* $\boxed{e}$ *kongkyekhan*]
[SBJ OBJ V-ADN]

b. *kica ka* [$\boxed{e}$ $\boxed{e}$ *kongkyekhan*]
[SBJ OBJ V-ADN]

Figure 8: Alternative syntactic roles for elements of the two prefix strings. Brackets indicate embedded clauses.

'The reporter met the senator who attacked (someone).'

(2) *kica ka* [$_{ORC}$ $\boxed{e}$ $\boxed{e}$ *kongkyekhan*] *uywon ul*
reporter NOM pro gap attack-ADN senator ACC
manassta.
meet-DECL

'The reporter met the senator whom (someone) attacked.'

(3) *kica ka* [$_{CC}$ $\boxed{e}$ $\boxed{e}$ *kongkyekhan*] *sasil ul*
reporter NOM pro pro attack-ADN fact ACC
alkoissta.
know-DECL

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

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 that 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 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 ul kongkyekhayssta.
senator ACC attack-DECL

'Someone attacked the senator.'

b. adjunct clause with a *pro*-subject

uywon ul kongkyekhayse kica ka
senator ACC attack-ADV reporter NOM
yumyenghaycyessta.
become-famous-DECL

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

c. complement clause with a *pro*-subject

uywon ul kongkyekhan sasil i palkhyecyessta.
senator ACC attack-ADN fact NOM is-revealed-DECL

'The fact that someone attacked the senator was revealed.'

d. subject relative clauses

uywon ul kongkyekhan kica ka
senator ACC attack-ADN reporter NOM
yumyenghaycyessta.
become-famous-DECL

'The reporter who attacked the senator became famous.'

e. matrix clause with a *pro*-object

kica ka kongkyekhayssta.
reporter NOM attack-DECL

'The reporter attacked someone.'

f. adjunct clause with a *pro*-object

kica ka kongkyekhayse uywon i
reporter NOM attack-ADV senator NOM
yumyenghaycyessta.
become-famous-DECL

'Because the reporter attacked someone/him, the senator became famous.'

g. complement clause with a *pro*-object

kica ka kongkyekhan sasil i palkhyecyessta.
reporter NOM attack-ADN fact NOM is-revealed-DECL

'The fact that the reporter attacked someone was revealed.'

h. object relative clauses

kica ka kongkyekhan uywon i
reporter NOM attack-ADN senator NOM
yumyenghaycyessta.
become-famous-DECL

'The senator whom the reporter attacked became famous.'

References

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, Massachusetts: MIT Press.

Frauenfelder, U., Segui, J., & Mehler, J. (1980). Monitoring around the relative clause. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 328 - 337.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, Massachusetts: MIT Press.

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

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643-672.

Hale, J. (under review). *What a rational parser would do.*

Han, N.-R., Ryu, S., Chae, S.-H., Yang, S., Lee, S., & Palmer, M. (2006). Korean treebank annotations version 2.0 [Computer software manual].

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior*, *20*(4), 417–430.

Ishizuka, T., Nakatani, K., & Gibson, E. (2006). *Processing Japanese relative clause in context.* Paper presented at the 19th Annual CUNY Conference on Human Sentence Processing.

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

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science*, *274*(5284), 114–116.

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

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language*, *30*(5), 580–602.

Kwon, N., Polinsky, M., & Kluender, R. (2006). Subject preference in Korean. In D. Baumer, D. Montero, & M. Scanlon (Eds.), *Proceedings of the 25th west coast conference on formal linguistics (WCCFL 25)* (p. 1-14). Somerville, MA: Cascadilla Press.

Lang, B. (1991). Towards a uniform formal framework for parsing. In *Current issues in parsing technology* (p. 153-171). Kluwer Academic Publishers.

Lee, C.-K. (2007). *Relative-clause processing in Korean adults: effects of constituent order and prosody*. Unpublished master's thesis, Rutgers University.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Lin, C.-J. C. (2008). The processing foundation of head-final relative clauses. *Language and Linguistics*, *9*, 813–838.

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)

O'Grady, W. (1997). *Syntactic development*. University of Chicago Press.

O'Grady, W., Lee, M., & Choo, M. (2003). A subject-object asymmetry in the acquisition of relative clauses in Korean as a second language. *Studies in Second Language Acquisition*, *25*(3), 433-448.

Stabler, E. (1997). Derivational minimalism. In C. Retoré (Ed.), *Logical aspects of computational linguistics.* Springer-Verlag.

Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language*, *52*(3), 452–473.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 119–161). Cambridge, Massachusetts: MIT Press.

Wilson, K., & Carroll, J. B. (1954). Applications of entropy measures to problems of sequential structure. In C. E. Osgood & T. A. Sebeok (Eds.), *Psycholinguistics: a survey of theory and research.* Indiana University Press.