Uncertainty and prediction in relativized structures across East Asian languages

Zhong Chen, Jiwon Yun, John Whitman, & John Hale (Cornell University)
zc77@cornell.edu

Relative clauses; Computational modeling; Chinese; Japanese; Korean

Incremental reading difficulties can be quantified using information-theoretic notions such as Entropy Reduction (ER, Hale, 2006). ER quantifies the amount of information contributed by a word in reducing structural uncertainties. This work extends ER to East Asian languages: Chinese, Japanese and Korean. This extension yields a quantitative viewpoint on ambiguities readers face when understanding prenominal relativized structures. It uncovers a range of language-specific factors that all pertain to the distribution on "unchosen" alternatives.

A Minimalist Grammar (Stabler, 1997) was written for each language. Weighting relevant construction types by treebank attestation counts allows us to estimate probabilistic "intersection" grammars conditioned on prefixes (Nederhof & Satta, 2008). Besides computing entropies, our system can also sample syntactic alternatives from intersection grammars to get an intuitive picture of how uncertainties are reduced during parsing.

Our modeling confirms the subject relative clause (SR) preference in Korean reported by Kwon et al. (2006) and further shows that this effect could emerge as early as the accusative/nominative marker in (1-2). This reflects, among other factors, a greater entropy reduction brought by sentence-initial nominative noun phrases.

Controversy has attended reports of a Chinese SR/OR asymmetry (Hsiao & Gibson, 2003; Lin & Bever, 2006). Our modeling derives an SR advantage in line with structural frequencies (SR 55% vs OR 45%). It also implicates headless RCs as a grammatical alternative whose existence makes processing easier at the head noun in SRs. A corpus study reveals that 14% of SRs have a null head whereas 31% of ORs are headless. This asymmetry suggests that an overt head is more predictable in SRs and less work needs to be done. All these predictions are derived from a grammar that covers various alternatives including *pro*-drops and *de* as a possessive marker.

The subject preference also holds in Japanese (Ishizuka, 2005; Ueno & Garnsey 2008). Kahraman et al. (2011), however, reported a puzzling inverse effect in Japanese clefts (5-6) that have the same word order as their relative clause counterparts. At the "-no-wa" marked embedded verb, object clefts are read faster than subject clefts. Our modeling technique derives a pattern consistent with this finding by tracking the frequency asymmetry between complement clauses and clefts. Upon reaching the topic marker "-wa", both constructions are still in play. But since complement clauses with object-*pro* are extremely rare, clefts become the more predictable structure.

In sum, examining contextualized syntactic alternatives shows how processing difficulty reflects the uncertainty associated with syntactic predictions. By using probabilistic grammars based on corpus counts, this methodology leverages a strong grammar-parser relationship.

(1) [e_i uywon -ul kongkyekhan] kica$_i$ -ka yumyenghaycyessta (Korean Subject Relatives)
senator ACC attack-ADN reporter NOM became-famous
'The reporter who attacked the senator became famous.'

(2) [kica -kae_i kongkyekhan] uywon$_i$ -i yumyenghaycyessta (Korean Object Relatives)
reporter NOM attack-ADN senator NOM became-famous
'The senator who the reporter attacked became famous.'

(3) [e_i yaoqing fuhao de] (guanyuan$_i$) da-le jizhe (Chinese Subject Relatives)
invite tycoon DE official hit reporter
'The official/Someone who invited the tycoon hit the reporter.'

(4) [fuhao yaoqing e_i de] (guanyuan$_i$) da-le jizhe (Chinese Object Relatives)
tycoon invite DE official hit reporter
'The official/Someone who the tycoon invited hit the reporter.'

(5) [e_i sobo -o kaihooshita -no] -wa shinseki$_i$ da (Japanese Subject Clefts)
grandma ACC nursed NO WA relative COP
'It was the relative who nursed the grandma.'

(6) [sobo -ga e_i kaihooshita -no] -wa shinseki$_i$ da (Japanese Object Clefts)
grandma NOM nursed NO WA relative COP
'It was the relative who the grandma nursed.'