Combining weight and discourse factors to predict relative clause extraposition in English

Elaine J. Francis
Purdue University

Laura A. Michaelis
University of Colorado at Boulder

In Relative Clause Extraposition (henceforth RCE), a subject-modifying clause occurs following the VP, as in (1a), rather than adjacent to the noun it modifies, as in (1b).

(1) a. New sets soon appeared that were able to receive all the TV channels. (ICE-GB corpus)
b. New sets that were able to receive all the TV channels soon appeared.

Structurally, RCE is unusual in that the modifying clause is not a sister to the N' head. Nevertheless, this construction occurs naturally in both formal and informal styles of spoken and written English.

Why should English allow (and sometimes prefer) a discontinuous structure as in (1a) when an adjacent ordering can express the same meaning? Most previous research has focused on discourse-based explanation: that RCE is preferred when the subject NP is focal and/or the VP is backgrounded (Huck & Na 1990; Kuno & Takami 2004; Rochemont & Culicover 1990; Takami 1999). This explanation accounts for the tendency of RCE tokens to contain unaccusative predicates and indefinite subject NPs (Lambrecht 1994). However, a recent corpus and experimental study by Francis (2010) offered a different motivation for RCE: RCE is preferred to the extent that the RC is longer than the VP, because late placement of "heavy" constituents facilitates language production and comprehension (Hawkins 2004, Wasow 2002). In the Francis study, both corpus and reading time data showed an advantage for RCE over non-RCE structure when the VP was short and the RC was long.

The present study, which combines a corpus analysis with two psycholinguistic experiments, shows that discourse and weight-based explanations are not mutually exclusive: both are important for predicting RCE. A binary logistic regression analysis of RCE and non-RCE sentences from the ICE-GB corpus (n = 345) reveals that ratio of VP length to RC length is the strongest predictor of RCE ($X^2(1) = 25.37$, $p < 0.01$), followed by definiteness ($X^2(1) = 11.78$, $p < 0.01$) and predicate type ($X^2(1) = 5.57$, $p = 0.018$). As shown in Figure 1, there is a strong preference for RCE when the weight ratio is less than 0.2 (i.e. the RC is more than five times longer than the VP), and a strong dispreference for RCE when the weight ratio is greater than 0.8. However, when the weight ratio is between 0.2 and 0.8, discourse-pragmatic factors become operative: the choice of RCE is determined primarily by definiteness and predicate type. For definite subject NPs, RCE is preferred only when the weight ratio is less than 0.2 (Figure 2), whereas for indefinite subject NPs occurring with a passive or unaccusative predicate, RCE is preferred with weight ratios up to 0.8 (Figure 3). Critically, discourse newness does not distinguish RCE from non-RCE sentences, since both types of tokens typically contain discourse-new subjects and predicates. Instead, RCE tokens are distinguished from non-RCE tokens by morphological form: RCE tokens typically contain a passive or unaccusative predicate and an indefinite or bare subject NP.

Two psycholinguistic experiments (in progress) follow up on these corpus findings. Both experiments use sentence materials which manipulate three factors: definiteness, RC length (5 words vs. 12 words), and VP length (2 words vs. 5 words). Predicate type is held constant, with a passive predicate in all experimental sentences. The first experiment, which consists of 64 experimental sentences and 96 filler sentences, measures structural preference in reading. Following Rosenbach (2005), participants are asked to choose which of two versions of a sentence (RCE vs. non-RCE) sounds more natural. The second experiment uses similar sentence materials to the first experiment to measure structural preference in production. Following Yamashita & Chang (2001), participants see sentence constituents randomly distributed on the computer screen, and they must formulate and speak a sentence using all of the parts. Preliminary results from thirty participants in the first experiment show very similar trends to the corpus data. As shown in Figure 4, RCE was preferred most often (74%) when the VP was short,

the RC was long, and the NP was indefinite, and least often when the VP was long, the RC was short, and the NP was definite (31%). A repeated measures ANOVA shows highly significant main effects for both RC length ($F = 31.85$, $p < 0.01$) and definiteness ($F = 58.09$, $p < 0.01$).

In conclusion, our corpus results and preliminary experimental results suggest that grammatical weight sets (soft) limits on RCE based on ease of processing, while discourse factors determine choice of RCE within these limits.

Figure 1: Percent extraposed by ratio of VP to RC length (ICE-GB corpus)

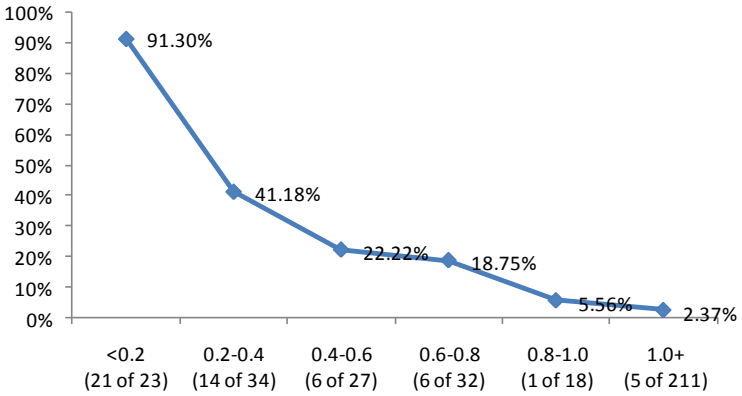

Figure 2: Percent extraposed by ratio of VP to RC length for definite subject NPs only (ICE-GB corpus)

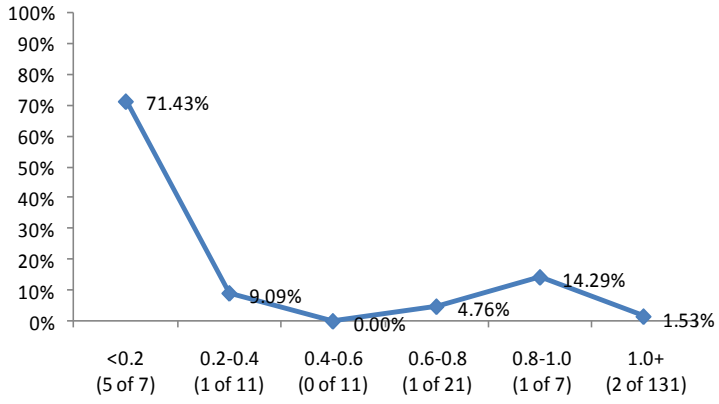

Figure 3: Percent extraposed by ratio of VP to RC length for indefinite subject NPs with passive or unaccusative predicate only (ICE-GB corpus)

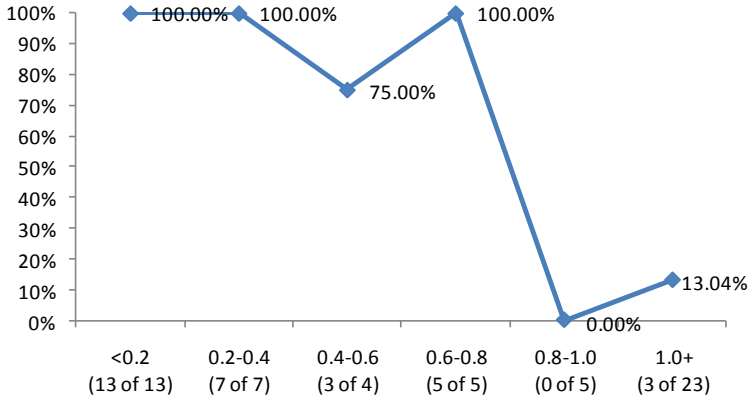

Figure 4: Percent extraposition responses in a structural preference task
(error bars represent standard error)