Understanding Complex Constructions: A Quantitative Corpus-Linguistic Approach to the Processing of English Relative Clauses

Daniel Wiechmann
daniel.wiechmann@uni-jena.de

1 Summary

The present work offers a corpus-based approach to the processing of complex constructions, specifically bi-clausal English relative clause constructions (RCC). At the most general level, the goal of the investigation is to help bridge the gap between linguistic and psycholinguistic research. Following a usage-based view on grammar (Langacker 2008), it is presumed that the processing demand of a linguistic structure is best portrayed as a function of a language user's prior experience with that structure. In the theoretical part of this paper, it is first argued that it is advantageous to assume a *constructionist* perspective on language (e.g Goldberg 2006), which discards a principled distinction between lexicon and grammar and instead holds that linguistic knowledge comprises of a large assembly of symbolic structures, which are termed *constructions*. These pairings of form and meaning may assume various degrees of specificity so as to incorporate linguistic material as concrete as a morpheme, but also material as abstract as a syntactic pattern. On this view, RCCs constitute highly abstract symbols, so called schematic constructions or *schemas*.

The constructionist conception of linguistic knowledge is then tied to an *exemplar-theoretic* view on language representation and processing (e.g. Bod 2009). In exemplar-based models of language, the processing of a linguistic structure is heavily dependent on the prior usage-frequency of that structure and, by implication, on the frequency of the pattern in the ambient language. Processing a linguistic structure then is classifying the structure relative to the inventory of structures already laid down in memory. As the mental representation of a linguistic structure is strengthened with each instance of usage and representational strength effects classification-time, frequent structures are easier to process. Patterns that occur with very high frequencies thus enjoy a privileged status in the cognitive system of a language user. In cognitive construction grammars, this

degree of representational strength is referred to as the degree of *entrenchment* of a structure.

While the entrenchment value of a construction is relatively easy to determine for simple constructions, where raw or relative frequencies in a representative sample of the language in question provide feasible approximations of entrenchment values, it becomes increasingly more difficult to estimate degrees of entrenchment of more complex construction types like RCCs. In the attempt to assess the entrenchment value of RCC-types in a statistically sound way, the study proposes a combination of association rule mining techniques (k-optimal pattern discovery, Webb and Zhang 2005) and pattern recognition techniques (hierarchical configural frequency analysis, von Eye and Pena 2004). An analysis of English RCCs of the kind envisaged here requires data sets that are both a) sufficiently large so that the analysis can reveal interesting effects and relationships and also b) ecologically valid in so far that the constructions under investigation reflect actual speakers solutions to functional, viz. communicative, pressures. To meet both of these requirements, a quantitative corpus-based approach to the issue was opted for and the study is based on a balanced corpus of contemporary British English (ICE-GB R2).

Since linguistic knowledge is highly structured and the processing mechanism is sensitive to that structure, the next step in the analysis targets the assessment of the structural relations among the detected patterns. In order to determine the underlying constructional network, RCC-types were related to each other on the basis of their degree of similarity. In the attempt to assign cognitively meaningful structures to the data, a hierarchical agglomerative clustering technique was employed that expresses similarity as distance in a Euclidean space and uses the neighbor-joining algorithm (Saitou and Nei 1987) to amalgamate groups of objects.

The next step in the analysis aims to show that the proposed model can predict human language processing behavior. To this end, the predictions of the corpus-derived constructional network of RCCs are compared to the experimental literature disclosing high degrees of compatibility with the more robust findings. The level of convergence between corpus-based and experimental findings is interpreted both as a success of the methodology employed and as a corroboration of the underlying theoretical framework.

Leaving the mechanistic plane, the study then turns to the question *why* certain patterns occur with above-chance probabilities in the ambient language in the first place. In other words, it is asked why speakers habituate themselves to particular RCC-patterns and not others? Here it is argued that certain patterns are dominant precisely because the discourse functions they encode are prominent in a given genre. In consequence, the detected patterns were related to research into information structure and discourse analysis. It is shown that the majority

of detected RCC-types can indeed be related to distinct discourse-functions that have been identified in the relevant literature, including *anchoring* new referents into the discourse (object relatives and transitive subject relatives), marking focus (cleft-like relatives), *channeling attention* (presentational relatives), or adding an *iconically shaped secondary predication* to the main clause predication (center embedded *-ed* participial RC).

The analysis departs from more traditional approaches to language processing, in so far as the latter typically assume that it is a set of intrinsic properties of a construction type (e.g. the complexity of that construction) that governs its processing difficulty. While architectural constraints from language production may govern certain linguistic choices under real-time pressure to some extent, it is argued that it is more felicitous to situate the effects of complexity at the social level, which embodies processes of conventionalization. The view presented here thus acknowledges an important connection between grammar and usage (Hawkins 2004) but suggests that the causal powers associated with the intrinsic properties of linguistic patterns first and foremost figure in the shaping of grammars over historical time, rather than in on-line processing.

2 References

Bod, Rens. 2009. From Exemplar to Grammar: A Probabilistic Analogy-based Model of Language Learning. *Cognitive Science*, Vol. 33(4).

Goldberg, Adele. 2006. *Constructions at work: The Nature of Generalisation in Language*. Oxford: OUP.

Hawkins, John. 2004. *Efficiency and Complexity in Grammars*. Oxford: OUP.

Langacker, Ronald. 2008. *Cognitive Grammar: A Basic Introduction*. New York: OUP.

Saitou, N., and Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406-425.

von Eye, A., and Pena, E. 2004. Configural Frequency Analysis: the Search for Extreme Cells. *Journal of Applied Statistics*, 31, 981-997.

Webb, G. I., and Zhang, S. 2005. k-Optimal-Rule-Discovery. *Data Mining and Knowledge Discovery*, 10(1). 39-79.