Elementary trees in RRGparbank

David Arps, Tatiana Bladier, Kilian Evang, Laura Kallmeyer, Robin Möllemann,
Rainer Osswald, Simon Petitjean
Heinrich Heine University Düsseldorf

RRGparbank is a parallel treebank under development which currently covers German, English, Farsi, French, and Russian, with Hungarian and Turkish to be added soon. Sentences in RRGparbank are annotated with RRG-compliant syntactic trees. At the moment, the main text corpus is George Orwell's novel "1984" and its translations into the respective languages. Syntactic annotations in RRGparbank are being created by automated parsing to Universal Dependencies [3], followed by an automated conversion to constituent trees and a subsequent manual correction step. The development of RRGparbank is driven by several goals: to provide a testbed for RRG as a formal theory of grammar; to provide empirical support for the study of cross-linguistic and language-specific properties of the languages under investigation; and to provide data sets for the development of NLP tools.

An important aspect of the project is the automatic decomposition of the annotated sentence trees into elementary trees. Elementary trees are here understood as lexically anchored syntactic trees which contain the full argument projection of their lexical anchors, but which are otherwise minimal (except for possible projections to the top). In particular, argument linking is already settled at the level of elementary trees and hence derived in another module of the grammar architecture. The decomposition is achieved without knowledge of the full set of available elementary trees. It is guided by general constraints on the form of elementary trees, by certain heuristics about head, complement and periphery identification, and by the assumption that syntactic composition in RRG can be formalized in terms of three operations over a finite set of elementary trees: simple substitution, sister adjunction and wrapping substitution [2, 4].

We extract elementary trees from RRGparbank using the top-down algorithm for automatic grammar extraction described in [1]. The decomposition is illustrated in Figure 1 by a French example that contains an inherently reflexive verb, a clitic and a fusion of a preposition and a definiteness operator. First results on the number of unanchored elementary trees per language are shown Table 1. Note that the numbers partly depend on the intermediate stage of the annotation process. The languages share about 30% of the elementary trees pairwise; 162 elementary trees are shared by all languages. This common subset contains a number of basic structures, such as frequent argument projections of nouns and verbs.

	German	English	Russian	French	Farsi
Annotated Sentences	5723	5452	4635	2000	1103
Annotated Tokens	85155	76918	50493	16886	8948
Elementary trees (ETs)	4208	3375	2954	1237	850
ETs occuring once in this language	2686	1981	1958	721	544
ETs occuring only in this language	2655	1847	1743	441	433

Table 1: Preliminary results on the automatic extraction of unanchored elementary trees. The numbers grow as the annotation of RRGparbank advances.

The tree decomposition, as well as the preceding annotation process, revealed several language-specific phenomena, such as complex nuclei in Farsi which combine verbs with NPs or PPs, clitics in French, or preposition stranding in English. Phenomena specific to German include the frequent occurrence of discontinuous nuclei, due to the separation of verbal particles, and the frequent use of extraposed relative clauses. A phenomenon specific to Russian is the use of deontic modal adverbials, which can scope over most constituents in a sentence instead of being restricted to the core. In the talk, we will show how these and other phenomena are reflected in the extracted elementary trees.

In future work we plan to use the extracted elementary trees for cross-lingual probabilistic parsing. Moreover, we plan to apply our data-driven approach for deriving linguistically adequate grammar sketches from the elementary trees which highlight both language-specific and cross-linguistically available syntactic configurations.

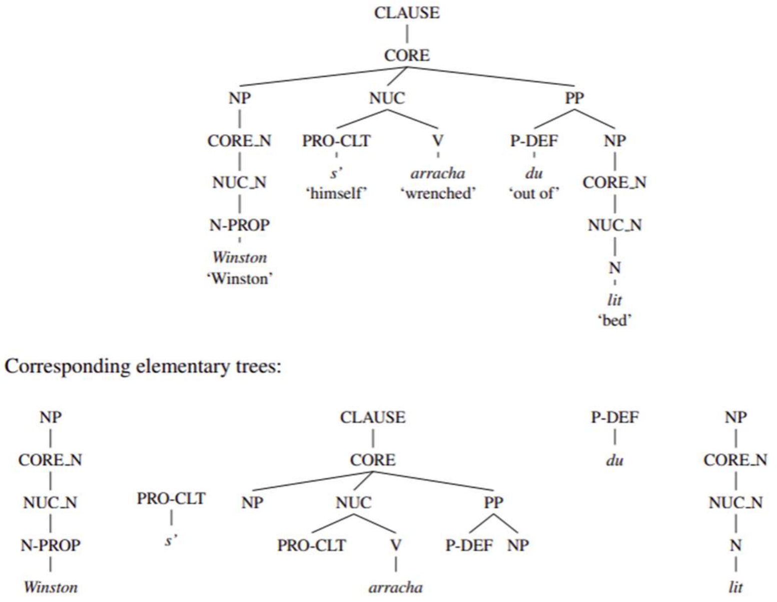

Figure 1: Decomposition of a sentence tree into lexically anchored elementary trees.

References:

[1] Tatiana Bladier et al. "Automatic Extraction of Tree-Wrapping Grammars for Multiple Languages". In: Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories. Dusseldorf, Germany: Association for Computational Linguistics, Oct. 2020, pp. 55–61. ¨ DOI: 10.18653/v1/2020.tlt1.5. URL: https://www.aclweb.org/anthology/2020.tlt-1.5. [2] Laura Kallmeyer, Rainer Osswald, and Robert D. Van Valin Jr. "Tree Wrapping for Role and Reference Grammar". In: Formal Grammar 2012/2013. Ed. by G. Morrill and M.-J. Nederhof. Vol. 8036. LNCS. Springer, 2013, pp. 175–190. [3] Joakim Nivre et al. "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection". English. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, May 2020, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.497. [4] Rainer Osswald and Laura Kallmeyer. "Towards a formalization of Role and Reference Grammar". In: Applying and Expanding Role and Reference Grammar. Ed. by Rolf Kailuweit, Eva Staudinger, and Lisann Kunkel. (NIHIN Studies). Freiburg: Albert-Ludwigs-Universitat, Universit ¨ atsbibliothek, 2018, pp. 355–378.