Towards a Role and Reference Grammar corpus for English

Christian Chiarcos, Christian Fäth & Monika Rind-Pawlowski
Applied Computational Linguistics, Goethe University Frankfurt, Germany

We describe an effort to create an openly available RRG treebank. At the time of writing, no publicly available RRG corpus seem to be in existence, although several initiatives to create rule-based and dictionary-based parsers have been around and are still being continued. Yet, without annotated data, no basis for their systematic evaluation is available, and neither, it would be possible to explore the usability and suitability, advantages or challenged that a semantically oriented formalism to syntax presents for purposes of state-of-the-art natural language processing or natural language understanding techniques which are largely based on machine learning. However, no machine learning without training data.

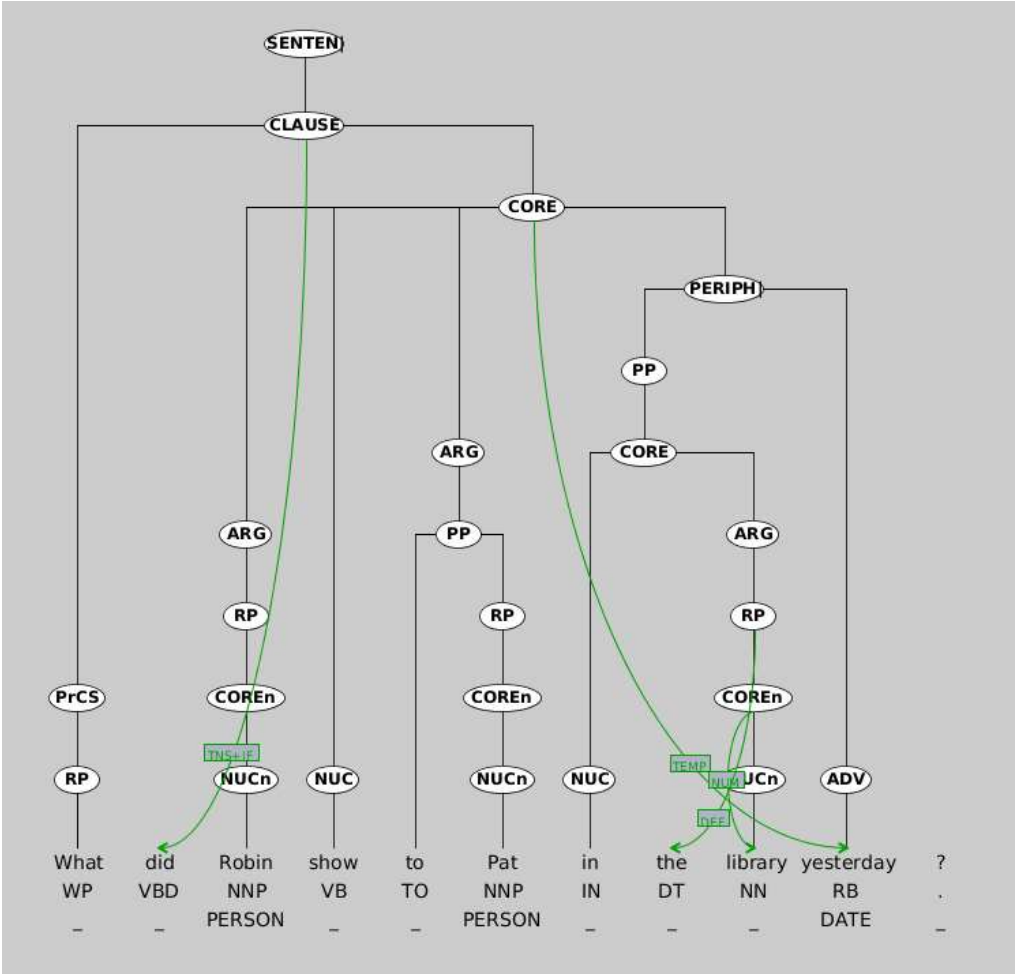
At the same time, the difficulties in creating RRG corpora have been identified, e.g., for a Quechua-Spanish corpus, whose creators, Rios and Göhring (2012), found that `the annotation process with RRG is too complex and error-prone' and thus shifted to simpler dependency-based syntax formalisms. As an alternative, we suggest deriving RRG annotations from existing, manually created annotations. Because of the specifics of RRG, and especially the great importance of semantics and verbal frames for its structure, no single resource is currently in existence from which am RRG corpus could be just derived. However, we argue that it is possible to derive valid RRG annotations from the intersection of existing annotation efforts in syntax on the one hand, and on verbal frames on the other hand.

We thus render the creation of an RRG corpus in terms of *annotation transformation* rather than *annotation*, and build on two cross-linguistically applied frameworks for syntax and semantics, respectively, the Universal Dependencies (UD) initiative (http://universaldependencies.org/) which currently provides syntactic annotations for more than 90 languages, and PropBank (PB, http://propbank.github.io/), a framework for the annotation of semantic frames of verbal (and nominal) predicates currently applied to 11 languages. Both resources are integrated at a deep conceptual level, where we

(1) derive CORE, arguments and periphery from semantic frames (PB),
(2) extrapolate the operator projection from morphosyntactic annotations (UD), applied to CORE and CLAUSE,
(3) derive nexus and juncture from shared semantic arguments (PB), operators (UD,PB), and clause linkage markers (UD),
(4) extrapolate remaining syntactic structures from UD.

Note that our approach is not a transformation, but a full decomposition and recomposition of various pieces of linguistic annotation according to RRG assumptions about their interaction. The resulting representation is thus richer than any annotation adopted as source. Also note that the underlying technology, CoNLL-RDF (Chiarcos and Fäth 2017), allows to consult external resources during the transformation, a functionality we use for disambiguating clause linkage markers and prepositions with verbal NUC, which are not distinguished in the underlying annotations.

A sample parse produced using automated annotations for Van Valin (2005, p.7, Fig.1.3) is illustrated below. Note that this visualization is produced with off-the-shelf tools for corpus querying developed by Lezius (2002), so it does not reflect the visual characteristics of Role and Reference Grammar but presents operator projection (green) and constituent projection (black/grey) in a compact, consolidated fashion. As for representing RRG analyses, we adopt a hybrid representation: For example, we keep RRG 1997 ARG labels as these can be reliably predicted from PB, but not automatically disambiguated without a designated RRG frame dictionary (which is not publicly available).



Over our GitHub repository (https://github.com/acoli-repo/rrg), this data is available for download and for use with the corpus tool TIGER Search (Lezius 2002). We provide two data sets under, a gold corpus comprised of all English examples from Van Valin and LaPolla (1997) and Van Valin (2008). This corpus was automatically annotated using the Stanford parser v. 1.6 by Manning et al. (2014), and the PropBank-compliant SRL system MATE by Björkelund et al. (2010), its annotations manually refined and both annotations then automatically transformed into an RRG representation.

The second dataset is a full transformation of the English Web Treebank, a corpus of 250,000 words with manual annotations for Universal Dependencies and PropBank semantic roles. This corpus was automatically converted and manually evaluated on the answers/dev section of the English Web Treebank. At the moment, we are awaiting copyright clearance for the original text, and, for the moment, provide the transformation scripts, instead, so that interested users can re-build the annotations locally.

The underlying technology is described with greater level of detail in a recent paper by Chiarcos and Fäth (2019). The interested reader may also want to compare a related effort recently described by Bladier et al. (2018), a second RRG corpus of English that was published shortly after our RRG Treebank. A key difference is that their data is created by converting from an existing syntax annotation, so that ambiguities that arise from RRG-specific information missing in the source annotation could not be resolved adequately. In comparison, we integrate information from two sources, semantic and syntactical, to arrive at a representation that is richer than either source annotation.

Our main interest in attending the International Conference on Role and Reference Grammar 2019 is to present our efforts and to discuss our preliminary achievements, but also

our limitations. In particular, we would like to explore possible directions for the validation of the generated treebank and discuss this with the wider RRG community. At the moment, we evaluate our efforts by matching the generated trees against known patterns originally devised for RRG *parsing*. However, it is evident that this form of evaluation is not only weak, but also limited to known patterns - whereas many of the constellations we produce are plausible (and in parts, they correspond to text book examples), they do not seem to conform to any of the previously published patterns. Furthermore, we see our efforts as complementary to rule-based and dictionary-based approaches for RRG parsing, with high potential for mutual synergies.

Acknowledgements
The research described in this paper was conducted in the context of the Early Carreer Research Group "Linked Open Dictionaries (LiODi)", funded by the German Ministry for Education and Research (BMBF).

References

Björkelund, A., B. Bohnet, L. Hafdell, and P. Nugues (2010). A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pp. 33-36. Association for Computational Linguistics.

Bladier, T, A. van Cranenburgh, K. Evang, L. Kallmeyer, R. Möllemann, R. Osswald (2018), RRGbank: A Role and Reference Grammar Corpus of Syntactic Structures Extracted from the Penn Treebank. in *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, pp. 5-16.

Chiarcos, C. and C. Fäth (2017). CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *Proceedings of the First Conference on Language, Data, and Knowledge - First International Conference (LDK 2017), Galway, Ireland, June 19-20, 2017*, pp. 74-88.

Chiarcos, C. and C. Fäth (2018). Graph-Based Annotation Engineering: Towards a Gold Corpus for Role and Reference Grammar. In *Proceedings of the 2nd Conference on Language, Data and Knowledge (LDK 2019), Leipzig, Germany, May*

Chiarcos, C. and N. Schenk (2018). The ACoLi CoNLL libraries: Beyond tab-separated values. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Lezius, W. (2002). TIGERSearch - ein Suchwerkzeug für Baumbanken. In *Proceedings of the 6. Konferenz zur Verarbeitung naturlicher Sprache (6th Conference on Natural Language Processing, KONVENS 2002), Saarbrücken, Germany*.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Rios, A. and A. Göhring (2012). A tree is a Baum is an árbol is a sach'a: Creating a trilingual treebank. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012*, pp. 1874-1879.

Van Valin, Jr., R. D. (2005). *Exploring the syntax-semantics interface*. Cambridge University Press.

Van Valin Jr, R. D. (2008). *Investigations of the Syntax Semantics Pragmatics Interface*, Volume 105. John Benjamins Publishing.

Van Valin, Jr., R. D. and R. J. LaPolla (1997). *Syntax: Structure, meaning, and function*. Cambridge University Press.