Creating RRG treebanks through semi-automatic conversion of annotated corpora

Tatiana Bladier, Kilian Evang, Laura Kallmeyer, Robin Möllemann & Rainer Osswald
Heinrich Heine University Düsseldorf

Introduction. Wide empirical coverage is an important issue for any grammatical theory. A strongly data-driven approach toward achieving this goal is the analysis and annotation of a sufficiently large collection of sentences which, ideally, captures all grammatical phenomena of a given language. An annotated corpus of this kind could then be used for testing the validity and the degree of coverage of hand-written grammar fragments as well as for the data-driven extension of such fragments. In this paper, we present ongoing work on developing treebanks for RRG, that is, corpora annotated with RRG compliant syntactic structures. Since large-scale syntactic annotation is a highly time-consuming task, our approach builds on existing annotations, which are transformed automatically into RRG structures. Our automatic conversion is developed on a small set of manually annotated sentences and combined with an additional manual correction cycle. We present two conversion algorithms: the first takes as input constituent structure annotations as used in the Penn Treebank (PTB) (Marcus, Marcinkiewicz, & Santorini, 1993), the second starts with dependency annotations in accordance with the Universal Dependency framework (UD) (Nivre et al., 2016).

Conversion procedure and annotation format. Our tree conversion process is iterative and error-driven, alternating between improving the conversion algorithm and comparing its output to manually validated RRG trees. We apply the conversion algorithm to bootstrap samples of new RRG trees, which are then checked and corrected by the annotators using the click/drag/drop-based web-interface we developed for the RRGbank (https://rrgbank.phil.hhu.de).

The usual notation of the RRG structures differs from the tree notation format typically used in treebanks, in that the RRG structures use an operator projection and represent periphery nodes and clause-linkage markers disconnected from the main constituency structure [cf. van Valin, 2005]. To avoid the discrepancy between the annotation formats, we adopted a notational variant for the RRGbank in which every RRG structure is represented as a single connected tree. We merge the operator projection (which is usually represented in the lower part of the

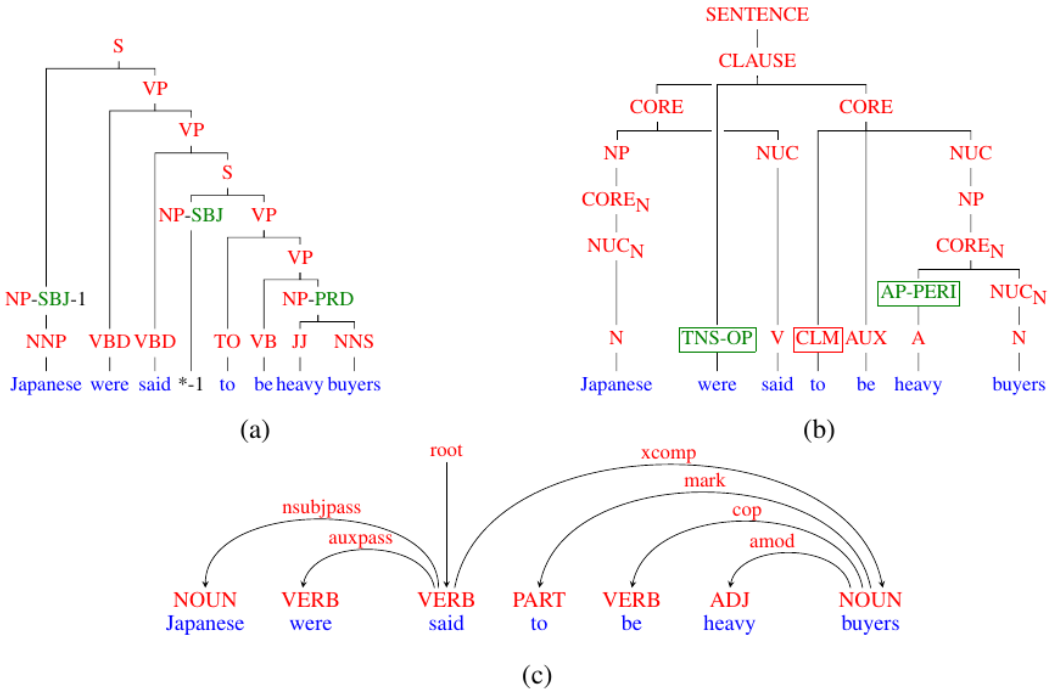

Figure 1: A tree from PTB (a) and PTB-UD (c) treebanks converted to an RRG structure (b).

RRG structures) with the constituency projection (see, for example, the position of the label for the tense operator *were* in Fig. 1b). We also attach peripheries (PERI) and clause-linkage markers (CLM) as daughters to the corresponding parent nodes preserving information contained in the original RRG structure. An example of our notational variant is shown in Fig. 1b, in which three boxes mark nodes represented differently in our notation.

Transformation of PTB to RRG structures. We chose the Wall Street Journal (WSJ) Sections of the Penn Treebank for conversion, which contains about 50.000 syntactically annotated sentences from WSJ articles. We randomly chose 500 sentences which contain different types of constructions and are no longer than 25 tokens. 100 sentences in this set were manually validated by at least two annotators with RRG expertise and the remaining 400 sentences have been manually corrected by one annotator. Our next goal is to increase the number of corrected sentences to 500 and add another 3000 shorter sentences (≤10 tokens). For transformation, we created a set of conversion rules, each applicable to constituents of a specific type. An example of a PTB tree transformed to an RRG structure is shown in Fig. 1b. We evaluated the performance of our conversion algorithm in terms of *completeness* and *correctness*. *Completeness* of the transformation was measured on the percentage of nodes in converted trees which have a label in the RRG label set. Since PTB and RRG share some labels (for example, PP, NP), the measured completeness amounted to 25.0% before conversion and 97.1% after conversion. We measured *correctness* by comparing converted trees with the manually annotated trees (i.e. our "gold trees"). The overall EVALB F1-score for the first 205 gold trees is 93.02.

Transformation of UDs to RRG structures. Universal Dependencies (Nivre et al., 2016) is a set of annotation guidelines for dependency trees. Like RRG, it emphasizes cross-linguistic applicability of all its structures (here: head-dependent arcs) and categories (here: functional labels of these arcs). The universal guidelines and the wide variety of existing treebanks for more than 70 languages make UD a promising starting point for creating a multilingual RRG resource where many conversion rules can be uniformly applied to corpora of different languages. An example UD tree is shown in Fig. 1c. We are currently in the process of developing an algorithm to convert UD trees to RRG structures by mapping each possible *local tree* (a node with its dependents and arc labels) to an RRG fragment and recursively converting the tree, starting from the root. The overall EVALB F1-score for the first 205 gold trees is currently 86.41.

Problematic and interesting cases during conversion. The main challenge for automatic conversion lies in phenomena for which RRG proposes different analyses, but which are not (explicitly) distinguished by PTB or UD. For example, PTB distinguishes prepositional arguments from prepositional adjuncts only in some cases that are marked with function labels such as -CLR. UD does not distinguish them at all. The control verb constructions are uniformly marked with traces in PTB and with the *xcomp* label in UD, while RRG requires two different analyses for these constructions depending on the nexus choice of the verb (i.e. co-subordination or coordination, see an example from PTB in Figure 2). In such cases, automatic conversion needs heuristic rules—e.g., based on lexical properties—to produce the correct RRG annotation. Comparing PTB and UD as input, we find that while UD trees are uniform across languages and less complex (e.g., they have no VP nodes), they are also slightly less informative, e.g., concerning adjunct vs. argument PPs and the nesting of coordinated NPs. Our converter still has to look at the original PTB tree to resolve these cases, illustrating that conversion from UD still needs some language-specific and treebank-specific rules. It remains to be seen which approach is ultimately superior. The conversion process also revealed several open questions in the RRG theory, which should be studied further, among which are for example RRG analysis of quantifier phrases such as "*more than 60%*". Finally, annotation errors exist in PTB, which need to be corrected manually.

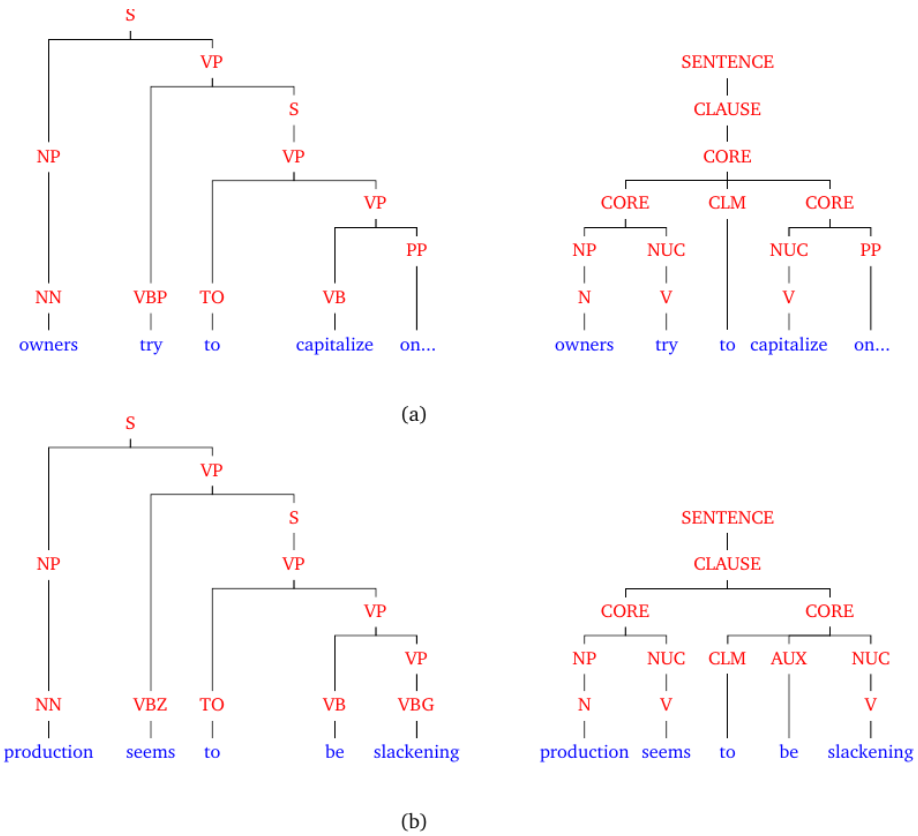

Figure 2: Core cosubordination (a) and core coordination (b) analysis in RRG for one construction in PTB.

Discussion and future work. In our future work, we plan to test and improve our conversion algorithm on other languages included in the UD corpus, starting with Russian and Tagalog. The ultimate goal is to be able to convert all corpora in the UD corpus to RRG corpora. We plan to make our converted corpora publicly available or available via the Linguistic Data Consortium depending on the original treebanks we use for the conversion.

References

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In K. Bontcheva & Z. Jingbo (Eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Stroudsburg, PA, USA: Association for Computational Linguistics.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.

Nivre, J., Marneffe, M.-C. de, Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., et al. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 1659–1666).

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