Automatic Extraction of Tree-Wrapping Grammars for German

Tatiana Bladier, Laura Kallmeyer
Heinrich Heine University Düsseldorf, {bladier, kallmeyer}@phil.hhu.de

Introduction. We present an extraction algorithm for Tree-Wrapping Grammar (TWG) for German from constituency treebanks. TWG [4, 5] is a tree-rewriting system inspired by Tree Adjoining Grammar (TAG) [3] that was developed for formalizing Role and Reference Grammar (RRG) [6, 7]. TWG aims, among others, at adequately representing long-distance dependencies (LDD). In this paper, we apply it to German data and inspect the LDD cases we find there. Furthermore, in order to deal with the free word order in German, we propose a slight extension of the TWG tree rewriting operations.

TWG. TWGs consist of elementary trees that are combined via a) *substitution* (replacing a leaf with a new tree), b) *sister-adjunction* (adding a new daughter to an internal node), and c) *wrapping substitution* (splitting the new tree at a d-edge (dominance link, notated as a dashed edge), filling a substitution node with the lower part and adding the upper part to the root of the target tree). We use lexicalized TWGs, i.e., each elementary tree has a lexical anchor. Fig. 1 gives an example.

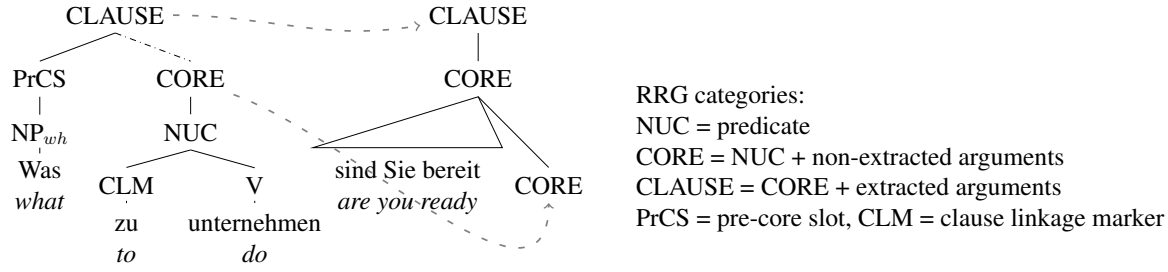

Figure 1: Wrapping substitution for wh-LDD

We extract a TWG from the German subcorpus of RRGparbank[1]. It turned out that for German a slight extension of *wrapping substitution* might be useful: We allow (i) for the upper part of the d-edge tree to also target internal nodes and (ii) for daughters of the upper d-edge node to be inserted in any place among daughters of the target node.

LDDs and wrapping in the extracted TWG We identified the following LDD cases[2] (the tokens which belong to LDDs are underlined, see also the visualized trees in Fig. 1 and Fig. 2):

- Long-distance wh-movement, for example *Was sind Sie bereit zu unternehmen ?*, see Fig. 1
- Complex predicates with non-canonical word order (*das Wasser ließ er ablaufen*, 'he let the water run off')).
- Relative clauses with LDDs (*den seine Haut auszudünsten schien*, 'that his skin seemed to exhale').
- Extraposed relative clauses (ERCs; *er hatte die Fotografie gesehen, die ihre Schuld widerlegte*, 'he had seen the photograph that disproved their guilt')

Wrapping steps for the second, third and fourth case are shown in Fig. 2. Nodes with the same NUC-ID (resp. CO-ID) features are actually a single multicomponent node with a discontinuous span in the original treebank tree. The PRED-ID feature in the treebank trees captures LDD information.

Comparison with previous work. Reported cases of LDDs in French and Dutch [2, 1] correlate largely with our findings in RRGparbank corpus data. Similarly to French and Dutch, the cases of LDDs in German data are relatively rare and affect less than 1 % of tokens in corpus. A peculiarity of German

[1] https://rrgparbank.phil.hhu.de/

[2] A case that we did not find in the treebank but that should nevertheless be considered is *scrambling*: ('*[. . .] dass es ihm der Junge zu reparieren zu versprechen bereit ist*', 'that the boy is ready to promise him to repair this.')

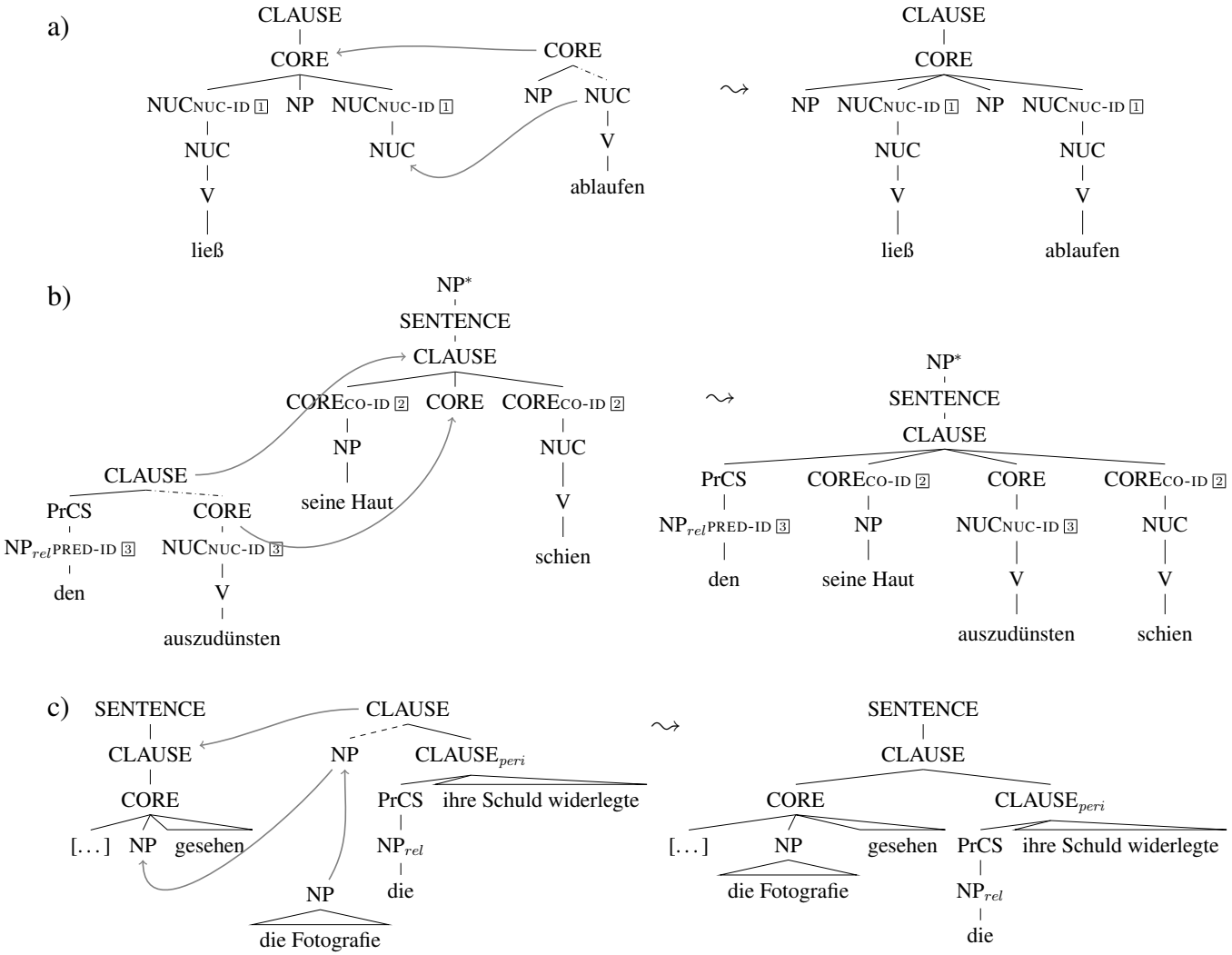

Figure 2: Wrapping: a) complex predicate; b) relative clause LDD (the latter adjoins via sister adjunction, indicated by the * on the root); c) extraposed relative clause.

is a large number of ERCs (105 of 118 LDD cases), which are not reported for French [2], but could be found in Dutch corpus Lassy Large [1]. Table 1 summarizes the findings in the present and previous work. The length of functional paths indicates the number of labels on the path between the dependent element in an LDD and its non-local governor node in the dependency analysis of the sentence [2].

	German (RRGparbank, this work)	**French** (FTB + Sequoia, [2])	**Dutch** (Lassy Large, [1])
# tokens total	87396	420169	>52M sentences
# tokens with LDD (%)	118 (0.14 %)	618 (0.15 %)	0.002– 0.01 % of sentences
# tokens with length of functional path = 2	109	513	–
# tokens with length of functional path >2	9	105	–

Table 1: Comparison of LDD cases with previous work for French and Dutch [2, 1].

TWG Extraction. We adapt the top-down algorithm from [8] for TAG. While substituting and sister-adjoining trees can be extracted as in [8], we developed a new algorithm to extract d-edge trees. As a preprocessing, we first remove crossing branches (see the multicomponent NUC and CORE in Fig. 2). Concerning LDDs, the parts of the LDD (indicated by PRED-ID and NUC-ID) are extracted within a single tree with a *d-edge*. More details will be given in the talk.

Outlook. We are currently performing TWG extraction and parsing for English, and by the time of the conference, we will further extend the annotation coverage for German and adapt our TWG extraction and parsing approach to it.

References

[1] Gosse Bouma. Corpus-evidence for true long-distance dependencies in dutch. *GRAMMAR AND CORPORA*, page 337, 2018.

[2] Marie Candito and Djamé Seddah. Effectively long-distance dependencies in French: Annotation and parsing evaluation. 2012.

[3] Aravind K Joshi and Yves Schabes. Tree-adjoining grammars. In *Handbook of formal languages*, pages 69–123. Springer, 1997.

[4] Laura Kallmeyer, Rainer Osswald, and Robert D. Van Valin, Jr. Tree Wrapping for Role and Reference Grammar. In G. Morrill and M.-J. Nederhof, editors, *Formal Grammar 2012/2013*, volume 8036 of *LNCS*, pages 175–190. Springer, 2013.

[5] Rainer Osswald and Laura Kallmeyer. Towards a formalization of Role and Reference Grammar. In Rolf Kailuweit, Lisann Künkel, and Eva Staudinger, editors, *Applying and Expanding Role and Reference Grammar.*, pages 355–378. Albert-Ludwigs-Universität, Universitätsbibliothek. [NIHIN studies], Freiburg, 2018.

[6] Robert D. Van Valin, Jr. and Randy LaPolla. *Syntax: Structure, meaning and function*. Cambridge University Press, 1997.

[7] Robert D. Van Valin Jr. *Exploring the syntax-semantics interface*. Cambridge University Press, 2005.

[8] Fei Xia. Extracting tree adjoining grammars from bracketed corpora. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99)*, pages 398–403, 1999.