Automatic extraction of Tree Wrapping Grammars from Discontinuous Constituent Treebanks

Tatiana Bladier, Laura Kallmeyer, Rainer Osswald, Jakub Waszczuk

TreeGraSP meeting #5
Düsseldorf, 11 November 2020

Tree Wrapping Grammar (TWG)

- Finite set of elementary trees, combined via:
  - → (simple) substitution,
  - → sister adjunction,
  - → wrapping substitution (Kallmeyer et al., 2013; Osswald and Kallmeyer, 2018).

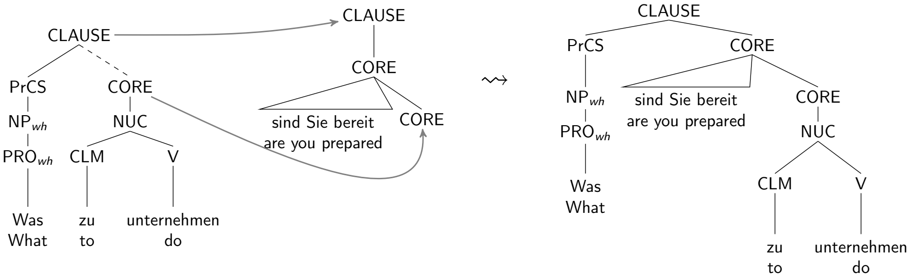


Figure 1: Wrapping substitution and a long distance dependency (LDD).

TWG: Extraposed Relative Clauses (ERCs)

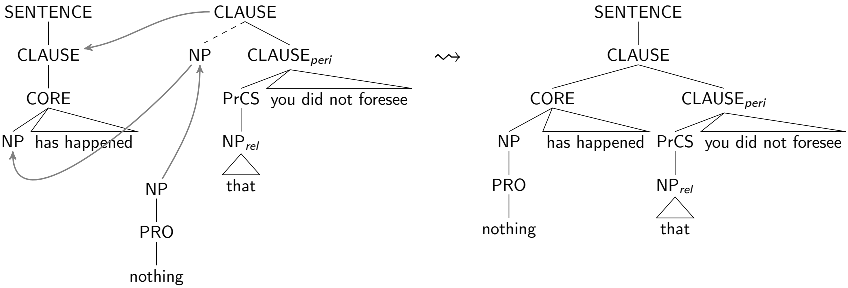
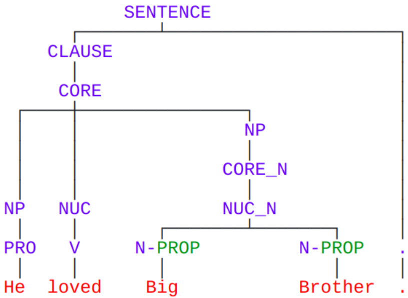


Figure 2: Extraposed Relative Clause (ERC) in TWG.

RRGparbank

- Multilingual corpus of RRG annotated sentences (RRG = Role and Reference Grammar, Van Valin and LaPolla 1997; Van Valin 2005)
- George Orwell's '1984' (≈ 6700 sentences) and translations into German, French, Russian, Farsi, Hungarian, Croatian
- Coverage of annotation so far: 81% English, 47% German, 12% French, 54% Russian, 15% Farsi
- rrgparbank.phil.hhu.de



TWG Extraction: initial and auxiliary trees

- TWG extraction algorithm based on Xia (1999) for TAG.
- Percolation tables for head and modifier distinction.

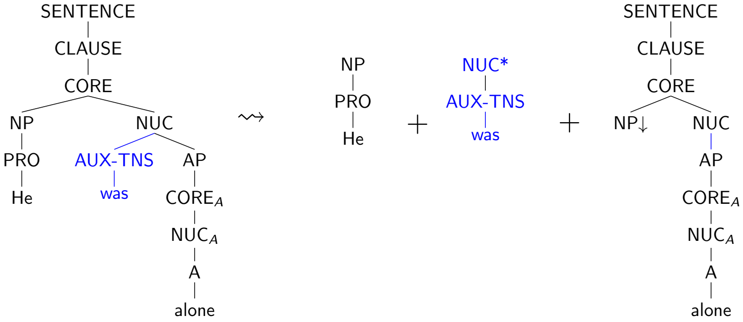


Figure 3: Extraction of initial and sister-adjoining trees.

TWG Extraction: d-edge trees for LDDs

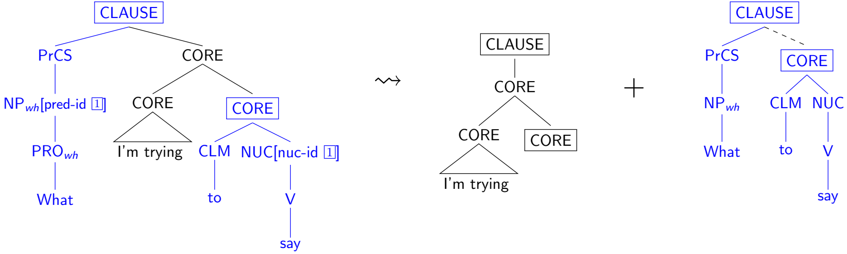


Figure 4: Extraction of a target tree and an elementary tree with a long distance dependency (LDD).

Extracted TWG Grammars

Parameters	English TWG	German TWG	French TWG	Russian TWG
# supertags	**3340**	**2591**	**947**	**2272**
# supertags occuring once	**1994**	**1689**	**584**	**1503**
# initial trees	1727	1490	483	1350
# sister-adjoining trees	1571	1031	431	898
# d-edge trees	**42**	**70**	**33**	**22**
Avg. sentence length	14.12	13.5	12.4	10.03
# sentences	5445	3062	851	3586
# Long-dist. dependenc. (LDDs)	**58**	**13**	**36**	**27**
# Extraposed rel. clauses (ERCs)	**8**	**110**	**4**	**0**

Table 1: Statistics on subcorpora and extracted grammars.

Similarity of extracted TWGs

Common supertags	**English TWG**	**German TWG**	**French TWG**	**Russian TWG**
English TWG	–	24.97 (834)	15.45 (516)	21.8 (728)
German TWG	32.19 (834)	–	15.51 (402)	24.9 (645)
French TWG	**54.49 (516)**	42.45 (402)	–	37.80 (358)
Russian TWG	32.04 (728)	28.4 (645)	15.76 (358)	–
Common supertags acr. languages	263			

Table 2: Ratio of common supertags across language pairs in percents and in numbers (in brackets).

Symbolic parsing with extracted grammars

	English TWG	German TWG	French TWG	Russian TWG
% exactly matching parses	81	79.07	78.86	80.68
# not parsed sentences	13	8	5	10

Table 3: Validation of extracted TWGs on symbolic parsing with TWG parser ParTAGe (Waszczuk, 2017; Bladier et al., 2020).

Problematic cases for TWG parsing

- Free-order placement of predicate arguments.

  Ja vot čto xoču skazat'.
  I here what want to.say
  'What I'm trying to say is this.'

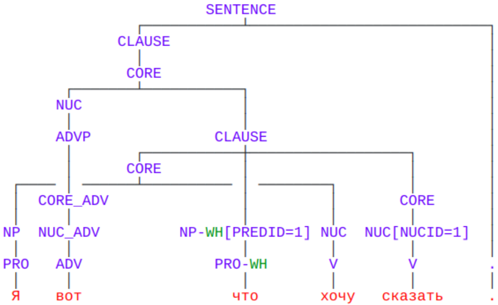


Figure 5: RRGparbank: interface example

Perspectives

Linguistic resources

- Corpus-based RRG grammars for different languages
- Dynamic treebanking for creating large RRG-annotated corpora
- Cross-linguistically valid “core” RRG grammar
- Cross-lingual proof of concept for TWG, in particular wrt. non-local dependencies

Parsing

- Wide-coverage probabilistic parsing with TWGs
- Multilingual TWG parsing

THANK YOU VERY MUCH
FOR YOUR ATTENTION!

References I

Bladier, T., Kallmeyer, L., and Waszczuk, J. (2020). Statistical Parsing of Tree Wrapping Grammars. Manuscript, accepted.

Kallmeyer, L., Osswald, R., and Van Valin, Jr., R. D. (2013). Tree Wrapping for Role and Reference Grammar. In Morrill, G. and Nederhof, M.-J., editors, *Formal Grammar 2012/2013*, volume 8036 of *LNCS*, pages 175–190. Springer.

Osswald, R. and Kallmeyer, L. (2018). Towards a formalization of Role and Reference Grammar. In Kailuweit, R., Künkel, L., and Staudinger, E., editors, *Applying and Expanding Role and Reference Grammar.*, pages 355–378. Albert-Ludwigs-Universität, Universitätsbibliothek. [NIHIN studies], Freiburg.

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

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

Waszczuk, J. (2017). *Leveraging MWEs in practical TAG parsing: towards the best of the two worlds*. PhD thesis.

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