Extraction of LTAG-based supertags from the French Treebank: challenges and possible solutions

Tatiana Bladier[1], Andreas van Cranenburgh[2], Laura Kallmeyer[1]
[1]Heinrich Heine University Düsseldorf
[2] University of Groningen
{bladier, kallmeyer}@phil.hhu.de, a.w.van.cranenburgh@rug.nl

Introduction. In the present paper we present an approach to automatically extract Lexicalized Tree Adjoining Grammars [LTAG; 6] from the French Treebank [FTB; 1], which can be used for the LTAG supertagging and parsing. We discuss the challenges encountered while extracting the grammars, propose our solutions, and evaluate our French LTAG grammars on the supertagging task.
LTAG is a linguistically and psychologically motivated grammar formalism. Productions in LTAGs support an *extended domain of locality*, which allows them to express linguistic generalizations that are not captured by typical statistical parsers based on context-free grammars or dependency parsing. Parsing with LTAGs can be facilitated through the intermediate step of *supertagging*—a task of assigning a sequence of LTAG tree templates (*supertags*) for a given sentence. Supertagging is referred to as being "almost parsing", since it takes much of syntactic disambiguation before applying a costly parsing algorithm [3]. Several supertagging approaches have been proposed for LTAGs [7, 2], with the most recent advances including approaches based on Recurrent Neural Networks (RNN) [9]. While supertagging experiments were reported for English [3, 9] and German [12], to our best knowledge, no research was reported on supertagging with French LTAGs.

Extraction of a feasible LTAG supertag lexicon. Neural supertagging with LTAGs is a sequence labeling problem. Thus, the performance of the supertagger on the classification task strongly depends on the size of the supertag lexicon—in case of a too large number of supertags, the supertagging system has to make a choice between a bigger number of options and thus suffers a drop of performance. Recent research on supertagging with automatically induced LTAG grammars for English and German [9, 12] shows that the manageable size of a supertag lexicon contains around 4000 distinct supertags with roughly a half of the supertags appearing only once (see Table 1).

LTAG Induction. We used the French Treebank for the grammar extraction due to its being the currently largest available and the most widely used resource for the French language. In order to extract an LTAG from the FTB, we applied the heuristic top-down procedure described by Xia [13]. For facilitation of the LTAG induction we carried out pre-processing steps described in Candito et al. [4] and Crabbé and Candito [5] including extension of the original POS tag set in FTB from 13 to 26 POS tags, undoing most multiword expressions with regular syntactic patterns and raising some complements (e.g. raising the PPs of the VPinf constituents). We experimented with the following LTAGs for French: including 13 or 26 POS tags, with or without compounds, including and excluding punctuation marks (see Table 2).

Challenges while extracting supertags from FTB. A big number of flat multi-word expressions (MWEs) in FTB leads to a large number of rather infrequent distinct extracted supertags. About 14 % of the word tokens in FTB belong to flat MWEs. After rewriting compounds with regular syntactic patterns, the number of MWEs is reduced to approximately 5 %. Extending the set of part-of-speech tags provided by FTB to more fine-grained 26 POS-tags theoretically helps the supertagger to better learn the dependencies between the supertags, however, a bigger number of POS tags leads to a higher number of supertags, which causes a drop of performance.

Left- and right-sister-adjunction. The tree structures in FTB are rather flat and allow any ordering of arguments and modifiers. In order to preserve these original flat structures as far as possible we decided against the traditional notion of adjunction in TAG which relies on nested structures and apply sister-adjunction; i.e., the root of a sister-adjoining tree can be attached as a daughter of any node of another tree with the same node label. Since a modifier can appear on the right or on the left side relative to the position of the constituent head, we distinguish between right- and left-sister-adjoining trees—marked with * on the left or the right side of the root label as shown in Figure 1.

Evaluation and future work. We evaluated our extracted grammars on experiments with an implemented supertagger similar to the one described in Kasai et al. [9] and Samih [10]. Our results show that the

extracted grammars get comparable results with the data for English and German LTAG. In our future work we plan to improve the supertagger and to use it for graph-based parsing. In particular, we aim at adapting the A*-based PARTAGE parser for LTAGs developed by Waszczuk [11] for parsing with extracted supertags. We also intend to add deep syntactic features and information on semantic roles to the supertags in order to test whether extracted LTAGs can be used for semantic role labeling.

Parameters	**French** (this work)	**German, reduced set** Kaeshammer [8]	**German, full set** Kaeshammer [8]	**English** Kasai et al. [9]
Supertags	5145	2516	3426	4727
Supertags occur. once	2693	1123	1562	2165
POS tags	13	53	53	36
Sentences	21550	28879	50000	44168
Avg. sentence length	31.34	17.51	17.71	appr. 20
Supertagging accuracy	**78.54**	85.91	**88.51**	**89.32**

Table 1: Comparison of LTAGs extracted from different treebanks

Extracted French LTAG	**# supertags**	**# supertags once**	**Supertagging accuracy**
13 POS, undone comp.	5145	2693	**78.54**
13 POS, with compounds	6847	3738	76.78
26 POS, with compounds	5831	3008	74.84
13 POS, undone comp., no punct. marks	5015	2557	74.44

Table 2: Supertagging experiments with different LTAGs extracted from the FTB.

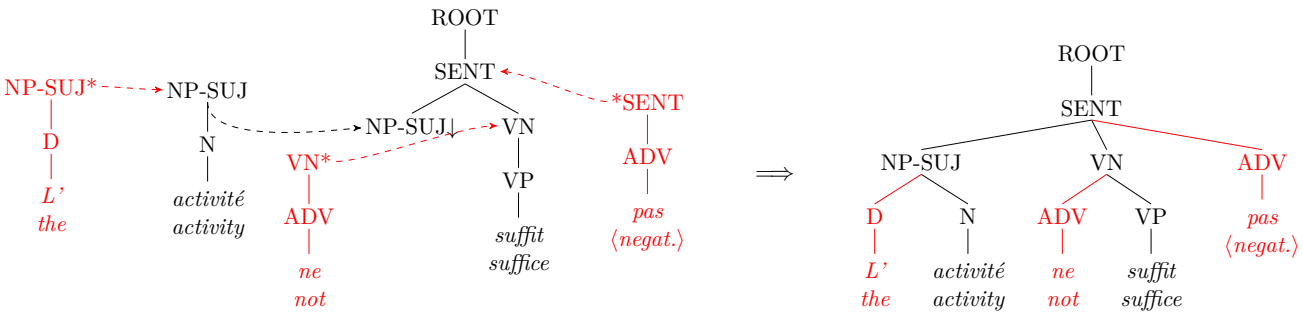

Figure 1: Left- and right-sister-adjunction for *L'activité ne suffit pas* ("The activity does not suffice")

References

[1] Anne Abeillé, Lionel Clément, and François Toussenel. Building a treebank for French. In *Treebanks*, pages 165–187. Springer, 2003.

[2] Jens Bäcker and Karin Harbusch. Hidden markov model-based supertagging in a user-initiative dialogue system. In *Proceedings of TAG+ 6*, pages 269–278, 2002.

[3] Srinivas Bangalore and Aravind K Joshi. Supertagging: An approach to almost parsing. *Computational linguistics*, 25 (2):237–265, 1999.

[4] Marie Candito, Benoît Crabbé, and Pascal Denis. Statistical french dependency parsing: treebank conversion and first results. In *Seventh International Conference on Language Resources and Evaluation-LREC 2010*, pages 1840–1847. European Language Resources Association (ELRA), 2010.

[5] Benoit Crabbé and Marie Candito. Expériences d'analyse syntaxique statistique du français. In *15ème conférence sur le Traitement Automatique des Langues Naturelles-TALN'08*, pages pp–44, 2008.

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

[7] Aravind K Joshi and Bangalore Srinivas. Disambiguation of super parts of speech (or supertags): Almost parsing. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 154–160. Association for Computational Linguistics, 1994.

[8] Miriam Kaeshammer. *A German treebank and lexicon for tree-adjoining grammars*. PhD thesis, Masters thesis, Universitat des Saarlandes, Saarlandes, Germany, 2012.

[9] Jungo Kasai, Bob Frank, Tom McCoy, Owen Rambow, and Alexis Nasr. Tag parsing with neural networks and vector representations of supertags. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1712–1722, 2017.

[10] Younes Samih. *Dialectal Arabic processing Using Deep Learning*. PhD thesis, Düsseldorf, Germany, 2017. URL http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:hbz:061-20180118-084822-9.

[11] Jakub Waszczuk. *Leveraging MWEs in practical TAG parsing: towards the best of the two world*. PhD thesis, 2017.

[12] Anika Westburg. Supertagging for german - an implementation based on tree adjoining grammars. Master's thesis, Heinrich Heine University of Düsseldorf, 2016.

[13] 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.