MarsaGram: an Excursion in the Forests of Parsing Trees

Philippe Blache, Stéphane Rauzy, Grégoire Montcheuil
Laboratoire Parole et Langage - UMR 7309
5 avenue Pasteur, 13100 Aix-en-Provence, France
firstname.lastname@lpl-aix.fr

Abstract

The question of how to compare languages and more generally the domain of linguistic typology, relies on the study of different linguistic properties or phenomena. Classically, such a comparison is done semi-manually, for example by extracting information from databases such as the *WALS*. However, it remains difficult to identify precisely regular parameters, available for different languages, that can be used as a basis towards modeling. We propose in this paper, focusing on the question of syntactic typology, a method for automatically extracting such parameters from treebanks, bringing them into a typology perspective. We present the method and the tools for inferring such information and navigating through the treebanks. The approach has been applied to 10 languages of the *Universal Dependencies Treebank*. We approach is evaluated by showing how automatic classification correlates with language families.

Keywords: Typology, Treebank, Syntactic properties, Grammar extraction

1. Introduction

Treebanks becoming available for many different languages (see for example (Nivre et al., 2015)), they are now important resources especially in a typology perspective. Associating syntactic information to large scale natural data offers both the extraction of regularities and the description of specific realizations of syntactic constructions. Furthermore, thanks to the use of a precise annotation guide, it becomes possible to extract automatically information and apply machine learning techniques to study the distribution of different phenomena.

However, the type of information classically encoded into a treebank remains at a high level of generality, that moreover very often remain implicit. For example, constituency-based treebanks encode an implicit phrase-structure grammar, that can be sometimes enriched with the annotation of the main syntactic relations. In this case, in order to render such information explicit, it is possible to automatically extract the grammar from the treebank and identify all the realizations of its rules. On top of this, it is also necessary to describe finer-grained information, such as government phenomena, linear order, cooccurrence, etc., which characterize better a language than the strict set of grammar's rules.

The same kind of questions arises in a typology perspective: comparing the grammars extracted from treebanks of different languages is not meaningful *per se*. On the other hand, typology very often focus only on specific information such as verb/arguments relation or head/modifiers orders.

We propose in this article an approach based on a specific representation of the syntactic information, in the perspective of characterizing languages in a typology point of view. This approach mainly relies on a new tool, *MarsaGram*[1], making it possible to automatically infer the typological information, independently of the formalism (constituency or dependency). We present in the following the type information that can be extracted and illustrate the approach with different treebanks.

2. Methods

2.1. Infering the context-free grammar of a constituency treebank

The extraction of a context-free grammar (CFG) from a constituency treebank is based on the classic method described in (Charniak, 1996): each internal node of the parsing tree is converted into a rule whose left-hand side (LHS) is the node's constituent tag and the right-hand side (RHS) is the sequence of children nodes' tags. The implicit grammar is formed by the set of such extracted rules. As an example, the figure 1 and 2 show the constituent tree for the French sentence "Elle a dix-sept ans." (*She is seventeen.*), and the corresponding CFG rules.

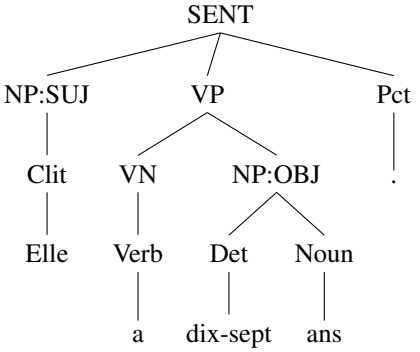

Figure 1: a small constituent tree

SENT	→	**NP:SUJ VP** *Pct*
NP:SUJ	→	*Clit*
VP	→	**VN NP:OBJ**
VN	→	*Verb*
NP:OBJ	→	*Det Noun*

Figure 2: inferred CFG rules

[1] Available on the Ortolang platform, hdl:11041/ortolang-000917

2.2. Filtering the rules

Our goal being to extract properties from the CFG grammar, which is highly dependent from the set of rules, the tool integrates different filters according the granularity of tags, the syntactic structures and the rule distribution:

Tag granularity Tags are made of the POS or the phrase type (e.g. *Noun*, **NP**) and the syntactic functions (e.g. **SUJ**). The *coarse grain* filter conserves only the first level of information and then groups syntactic units independently from their function.

None elements Some formalisms make use of components without projection in order to represent some linguistics phenomenons (such as ellipsis) or indicate relations between constituents. The *none* filter suppress these empty elements and their traces in the tree.

Coordinations Coordination constructions are frequent phenomena, that multiply the set of realized rules without providing much information on the relations between constituents. The *coordination* filter reduces expressions that match certain prototypical coordination patterns.

Frequency A large number of rules (in other words local trees) appear rarely in the treebank, which can introduce noise. The *frequency* filter remove rules based on a minimal number of occurrences, or their ranks or frequencies among the rules of same LHS.

Experiment

We use in this experiment 4 constituency treebanks, three of them being in the Penn treebank framework: the *Penn Treebank* itself (Marcus et al., 1995) itself, the *Chinese Treebank* (Xue et al., 2010) and the *Arabic Treebank* (Maamouri et al., 2010a; Maamouri et al., 2011; Maamouri et al., 2010b) all of them distributed by LDC. We also use the *Modified French Treebank* (Schluter and van Genabith, 2007) a subset of the *French Treebank* (Abeillé et al., 2003) developed and distributed by LLF. The table 1 shows the extracted grammars sizes, when applying different filters.

2.3. Adaptation to dependency treebanks

We apply a similar approach to dependency treebanks: internal nodes (LHS of the rules) are now the head category, and children nodes (RHS) are the list of dependents, respecting their projection order, plus an extra node with the symbol "*" indicating the head projection. The figures 3 and 4 show the dependency tree for the same French sentence of figure 1 and the CFG rules extracted.

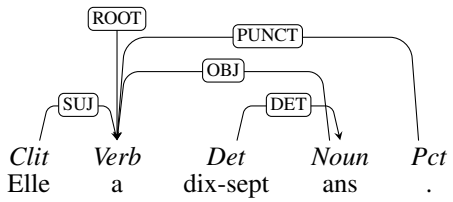

Figure 3: a small dependency tree

Verb:ROOT	→	*Clit:SUJ*	*	*Noun:OBJ*
		Pct:PUNCT		
Noun:OBJ	→	*Det:DET* *		

Figure 4: infered CFG rules

2.4. Inducing the syntactic properties

It remains difficult to explore or compare the syntactic characteristics of various languages using the complete grammar (independently of the formalism, constituents or dependencies). On the other hand, it is possible to compare some specific properties, in accordance with established practices in typology. For example, a classical typology consists in studying the verb/arguments relations and their linearity. We propose to induce from the treebanks some properties used to settle typologies. This properties are identified in the frame of *Property Grammar*.

Our tool focuses on 4 types of properties governing syntactic components, as described in (Blache and Rauzy, 2012).

Linearity : two components (A,B) have a linearity relation when the occurrence order of these two components is always the same.

Requirement : two components (A,B) have a requirement relation when the presence of one requires the presence of the other.

Exclusion : two components (A,B) have an exclusion relation when they do not occur together.

Unicity : a component A has an unicity property if it never occurs several times in the RHS of the rules with a same LHS.

We note a property as a *4-tuple* $p =< C, rel, A, B >$ where C is the context (i.e. the LHS component), rel one of the relations ($precede, require, exclude, unicity$) and A and B are the two components ($B = A$ in the case of $unicity$).

Validating and violating subsets

In order to infer these properties, for each distinct LHS (syntactic unit or head), we first compile the set of components (i.e. the elements that occur in a RHS). In a second stage, for each pair of components (A, B), we separate the rules where A and/or B occur (together or not) according to whether they satisfy or violate a property as show on the table 2. Starting from these subsets, it becomes possible to consider only the properties that are always satisfied (i.e without any violating rules). It is also possible to relax this constraint and make use of weighted properties, distinguishing between *strong* (frequent and (almost) always satisfied) and *weak*er (less frequent and/or more often violated) properties.

Weigthing the properties

A first weight w_0 (equation (1)) can be directly obtained by calculating the ratio of occurrences of the validating rules to the sum of both subsets — the properties satisfied in all cases corresponding then to $w_0 = 1$. However, if w_0 allows a first filtering of the properties, it does not provide any information about their actual weight. Suppose that two components (A, B) occur very rarely together: it is highly

Treebank	#trees	#rules					
		fine grain	coarse grain				
				reduc(none)	reduc(coord)	reduc(none) reduc(coord)	reduc(none) reduc(coord) $occ(r) \geq 2$
PTB (en)	49.786	31.291	17.446	17.120	16.417	16.085	6.867
ChTB (zh)	51.447	28.128	13.941	14.082	13.509	13.652	5.232
ATB (ar)	23.488	30.810	16.816	16.669	16.266	16.116	6.945
MFT (fr)	3.774	6.244	3.363	-	3.300	-	1.391

Table 1: some context-free grammars sizes

property p	$r \in Validating(p)$	$r \in Violating(p)$
$< C, precede, A, B >$	A and B occur in the RHS of r **and** all occurrences of A are before the first B	A and B occur in the RHS of r **and** at least one occurrence of A is after a B
$< C, require, A, B >$	A and B occur in the RHS of r	A occurs, but B no.
$< C, exclude, A, B >$	Only A or only B occur in the RHS of r	Both A and B occur
$< C, unicity, A, A >$	Only one A occurs in the RHS of r	Various A in the RHS of r

Table 2: validating and violating rules

probable that one of their linearity relations has a w_0 of 1 (or close to 1); nevertheless it would be more relevant to consider the exclusion property even if its w_0 is lower. We introduce then a second weight w_1 (equation (2)) that balances w_0 with the frequency of validating rules in relation to the entire set of rules (with the same LHS i.e. *Rules(C)*).

$$w_0 = \frac{Occ(Validating(p))}{Occ(Validating(p)) + Occ(Violating(p))} \quad (1)$$

$$w_1 = w_0 \frac{Occ(Validating(p))}{\sigma_{Occ}} \quad (2)$$

Where: $\sigma_{Occ} = \sum_{r \in Rules(C)} Occ(r)$

Experiment

This experiment relies on the first version of the *Universal Dependency Treebank* (Nivre et al., 2015), a family of dependency treebanks for 10 languages (Czech, German, English, Spanish, Finnish, French, Irish, Hungarian, Italian and Swedish) using an unique tagset : 17 Part-Of-Speech tags based on (Petrov et al., 2012), and 40 standardized dependencies relations described in (de Marneffe et al., 2014). As all languages use a common tagset, we can compare the properties extracted from each treebank and define a similarity between two languages based on the proportion of common properties. We build a hierarchical clustering of the 10 languages of the *Universal Dependency Treebank* (using the $hclust$ function of *R*). The figure 5 shows the resulting dendrograms either using all types of properties or only the linearity properties ($precede$). If the first clustering, using all properties does not allow to bring out clear aggregations. The second one, restricted to linearity properties, is more convincing: we find in a same branch the romance languages (es, it and fr), two of the germanic languages (de, sv) — English is grouped with the romance languages before joining the germanic branch — or the two uralic languages (fi, hu) (see table 3 for language families and features).

3. Visualizing the data

MarsaGram is also a tool to visually explore treebanks. The program genarates a set of dynamic html pages to navigate, for each LHS, through the elements composing the RHS, the extracted properties and the list of retained and filtered rules. The figure 6 shows a general view of the interface, divided in three part: an index on the left side, the data information on the main part and a visualization of treebank's occurrences in the lower part. The index lists all symbols of the treebank with their frequency and, for the symbols that are the LHS of some rules, the number of associated rules and extracted properties. A link on each LHS symbol loads the corresponding information on the main part of the windows.

The main part offers three combinable "axes" of exploration: the list of symbols composing the RHS part of the rules, the extracted properties and the retained rules. Three tabs allow to switch between these axes as a starting point. Then we can progress in the exploration, combining the others axes as show the figures 7 to 10. Starting with the list of properties (figures 7, for the verbal phrase (VP), filtered to show only the *precede* relation), a click on a property opens a sub-table listing the various rules indexes that are related to the property, as shows figure 8 (property $< VP, precede, VDB, NP >$, where *VBD* is a verb at the past tense and NP a noun phrase). The table shows both the rules validating the property (here the *precede* row) and the violating ones (here the *follow* row, with a dark orange background). When the mouse passes over a rule index, the

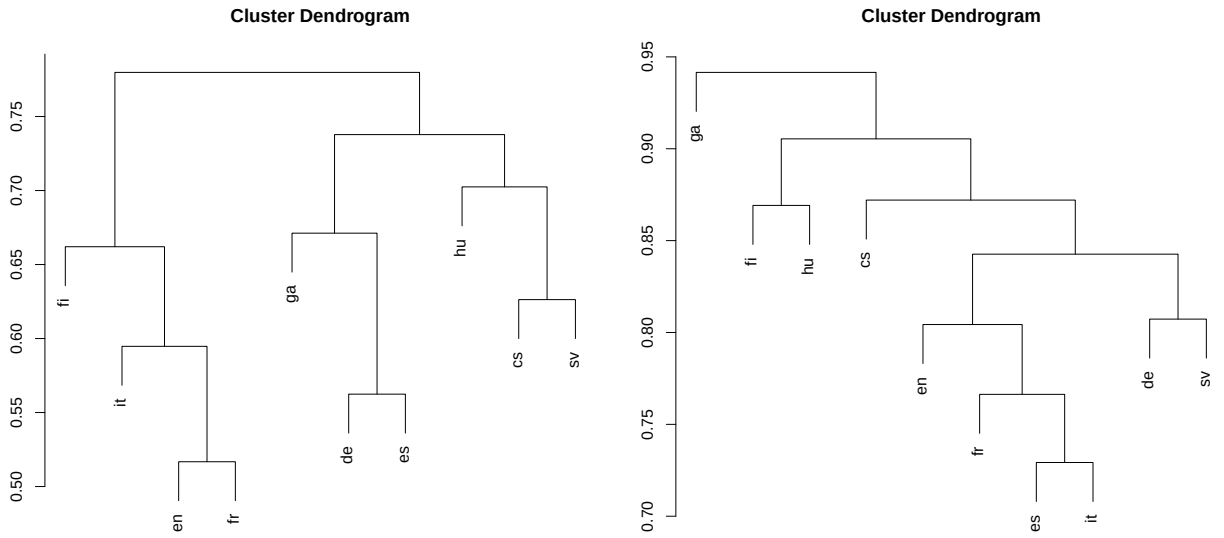

Figure 5: Comparing languages : all properties(left) and linearity properties(right)

Code	Language	Family	Genre	typologic features
cs	Czech	Indo-European	Slavic	SVO[2], stress-timed, free word order
de	German	Indo-European	Germanic	V2 and SOV, inflectional, accusative, stress-timed, dynamic accent
en	English	Indo-European	Germanic	SVO, inflectional, accusative, stress-timed, dynamic accent
sv	Swedish	Indo-European	Germanic	SVO, inflectional, accusative, stress-timed, pitch accent
es	Spanish	Indo-European	Romance	SVO, syllabic
fr	French	Indo-European	Romance	SVO, inflectional, accusative, syllabic
it	Italian	Indo-European	Romance	SVO, syllabic
ga	Irish	Indo-European	Celtic	inflectional, accusative, VSO, stress-timed, dynamic accent
fi	Finnish	Uralic	Finnic	SVO, free word order
hu	Hungarian	Uralic	Ugric	SOV, free word order, agglutinative, accusative

Table 3: some languages features

rule appears as a tooltip, and by a click it is selected to appear on a sub-table showing rules and some data (figure 9 : occurrences, number of validated and violated properties). From these rules data table, a click opens a sub-table (figure 10) with more details on the rule: its several variants when some reductions are used (none elements, coordination) and, for each variant, the number of occurrences and their indexes in the treebank, with a link that shows the actual occurrence in the lower part of the window.
When filters are used, another tab allows to inspect the list of filtered rules and access in a similar way to their actual occurrences.

4. Conclusion

We have presented in this paper the method used to infer properties from treebanks, independently of the formalism (constituency or dependency). This method integrates different filtering options making it possible to tune the properties extraction. An application of this method to the *Universal Dependencies Treebank* has shown the validity of the approach in a typology perspective: a hierarchical clustering based on a subset of properties led to the reconstitution of language families. The system we have developed in this perspective, *MarsaGram*, is freely available and integrates as a side effect a possibility to use it as a treebank browser.

5. Acknowledgments

This work has benefited from the French State support via its "Investissements d'avenir" programs ORTOLANG (ANR-11-EQPX-0032), BLRI (ANR-11-LABX-0036) and A*MIDEX (ANR-11-IDEX-0001-02).

6. Bibliographical References

Blache, P. and Rauzy, S. (2012). Enrichissement du FTB : un treebank hybride constituants/propriétés (Enriching the French Treebank with Properties) [in french]. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 307–320, Grenoble, France, June. ATALA/AFCP.

[2]SVO, SOV or VSO are the relative order of the Sujet, Verbe and Objet; and V2 stand for the verbe in second position.

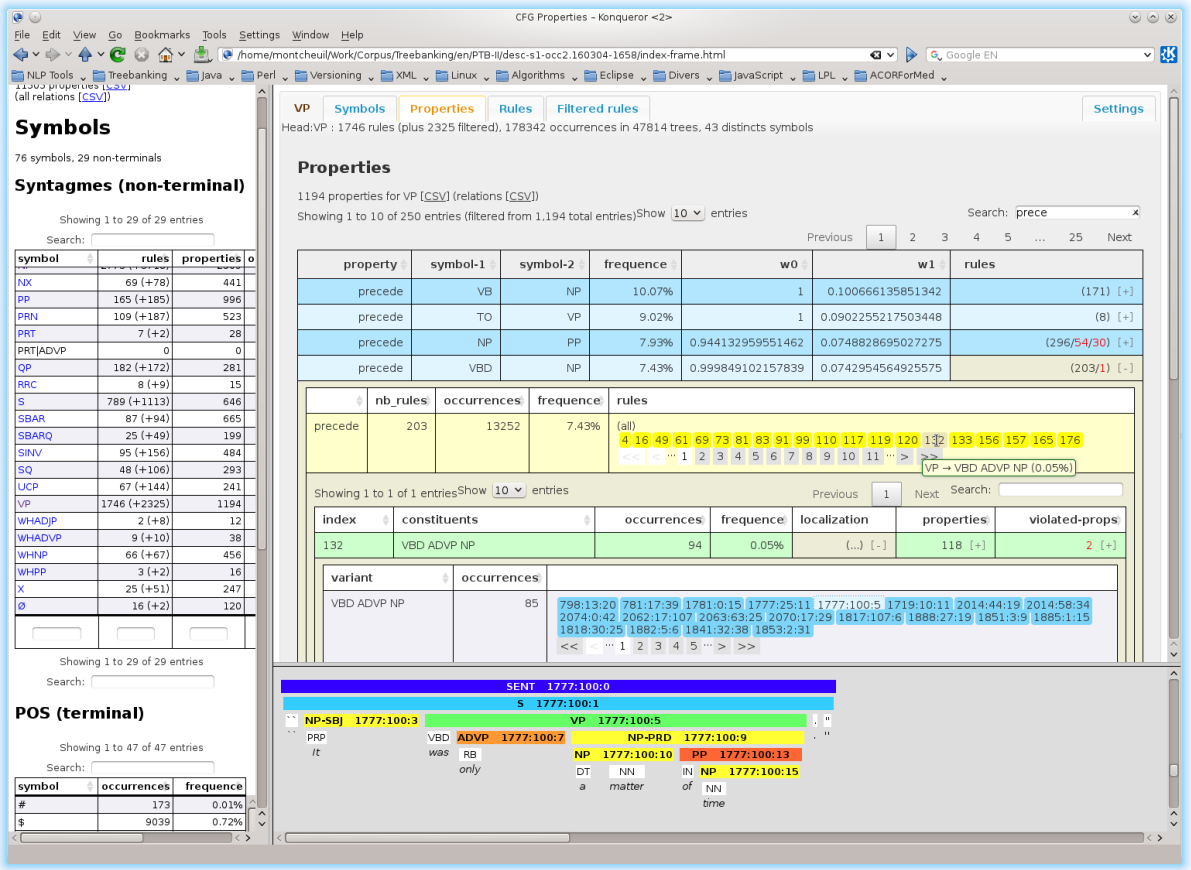

Figure 6: MarsaGram interface

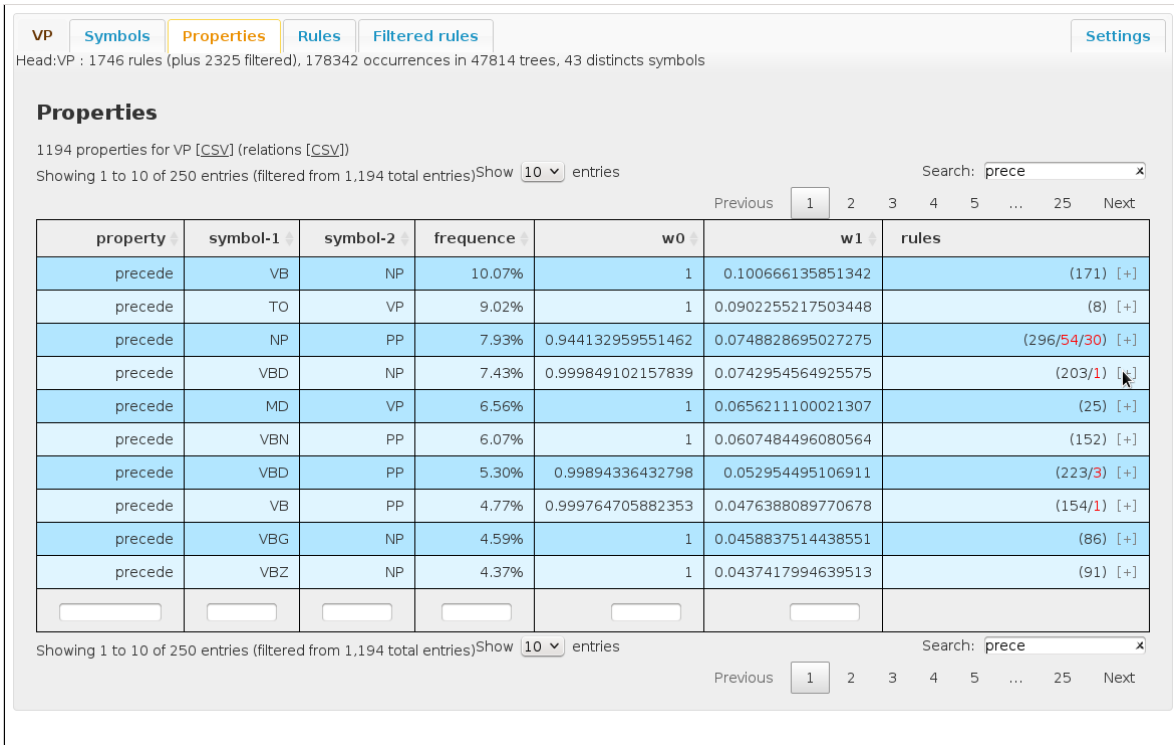

Figure 7: exploring properties

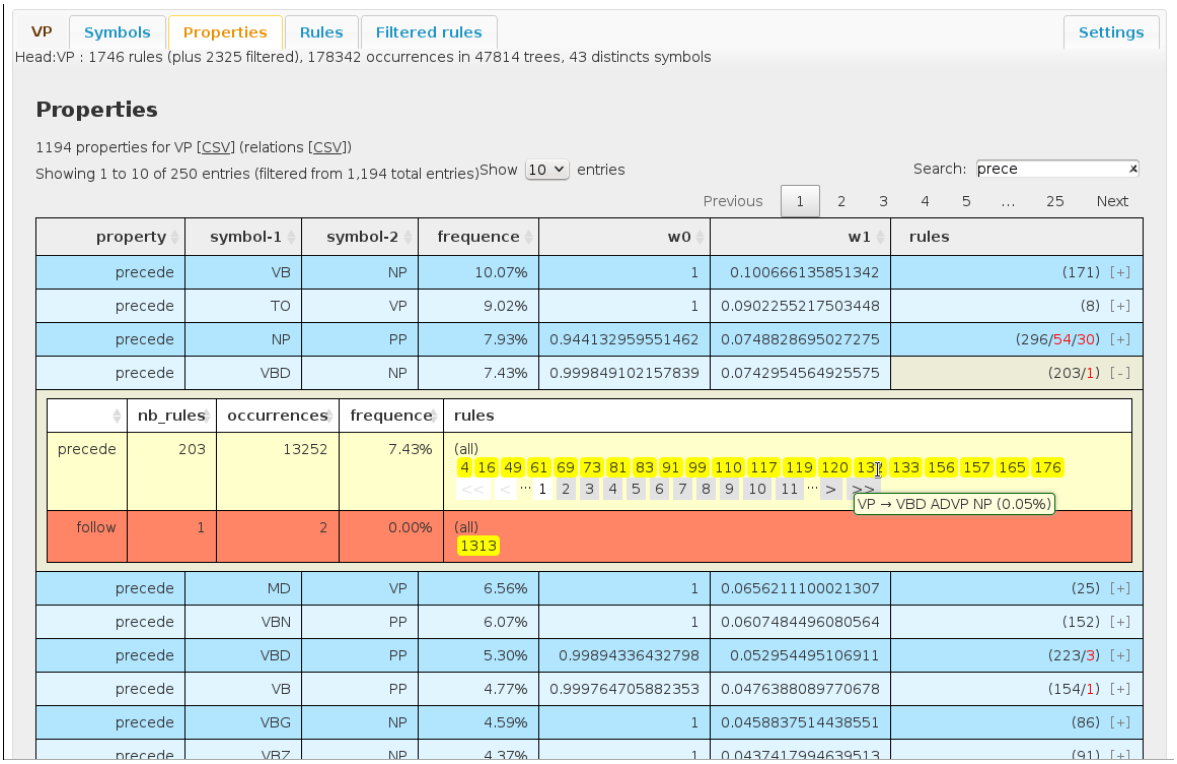

Figure 8: properties' rules

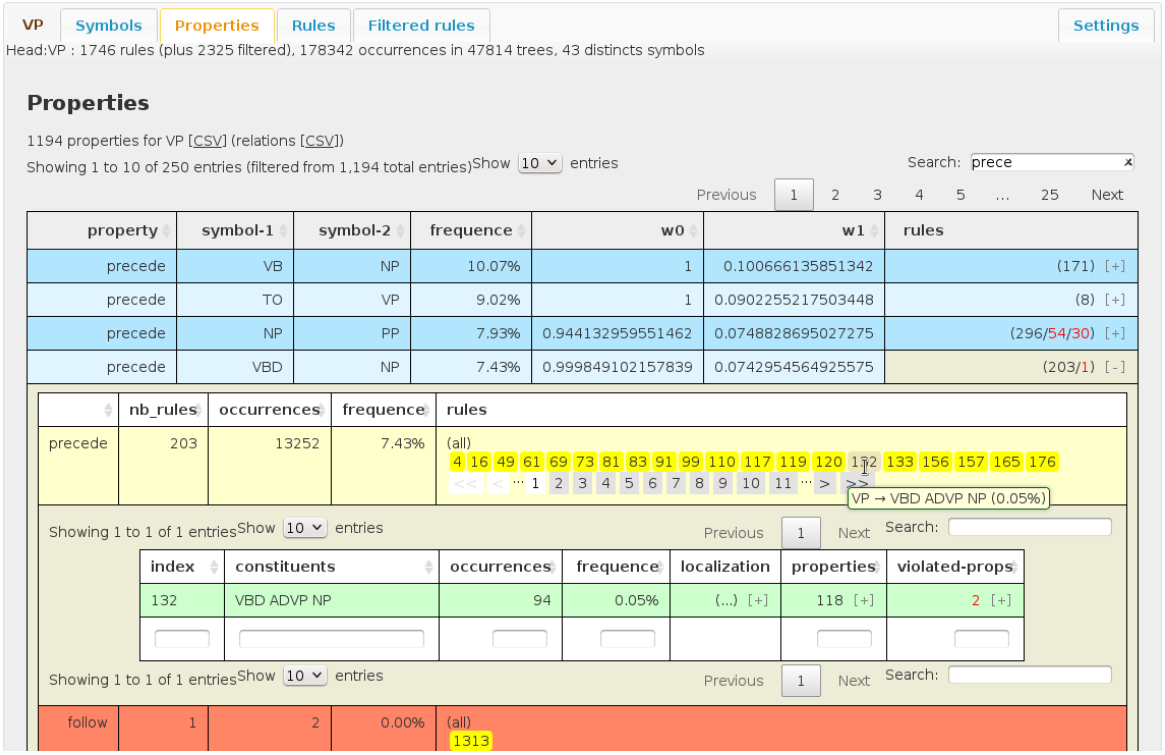

Figure 9: a specific rule ($VP \rightarrow VDB\ ADVP\ NP$)

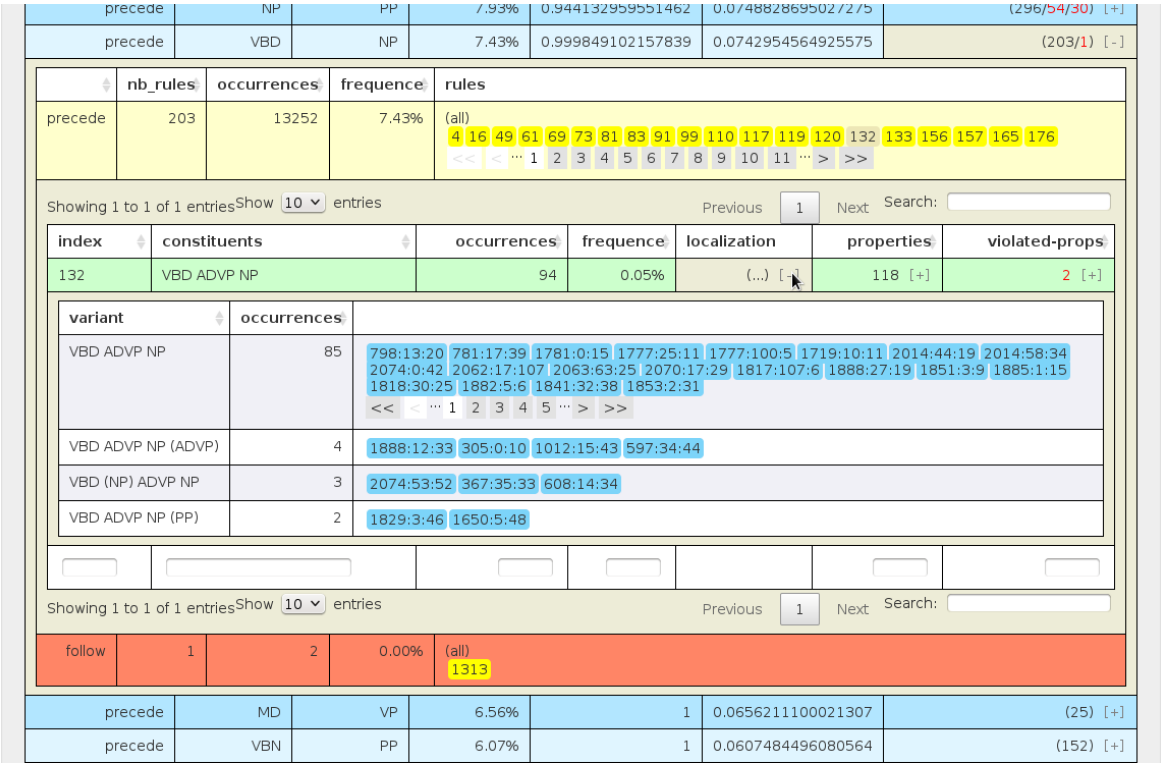

Figure 10: rule's variants and localisation

Charniak, E. (1996). Tree-bank grammars. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1031–1036.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and D. Manning, C. (2014). Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

7. Language Resource References

Abeillé, A., Clément, L., and Toussenel, F. (2003). *Corpus arboré pour le français / French Treebank*. Laboratoire de Linguistique Formelle. URL: `http://www.llf.cnrs.fr/fr/Gens/Abeille/French-Treebank-fr.php`.

Maamouri, M., Bies, A., Kulick, S., Gaddeche, F., Mekki, W., Krouna, S., Bouziri, B., and Zaghouani, W. (2010a). *Arabic Treebank:Part 1 v4.1*. Linguistic Data Consortium, Arabic Treebank, 4.1. ISLRN 512-715-458-848-0.

Maamouri, M., Bies, A., Kulick, S., Krouna, S., Gaddeche, F., and Zaghouani, W. (2010b). *Arabic Treebank:Part 3 v3.2*. Linguistic Data Consortium, Arabic Treebank, 3.2. ISLRN 770-467-034-042-0.

Maamouri, M., Bies, A., Kulick, S., Gaddeche, F., Mekki, W., Krouna, S., Bouziri, B., and Zaghouani, W. (2011). *Arabic Treebank:Part 2 v3.1*. Linguistic Data Consortium, Arabic Treebank, 3.1. ISLRN 758-179-408-820-5.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1995). *Penn Treebank 2*. Linguistic Data Consortium, Penn Treebank, 2. ISLRN 650-146-755-602-3.

Nivre, J., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goldberg, Y., Hajič, J., Kanerva, J., Laippala, V., Lenci, A., Lynn, T., Manning, C., McDonald, R., Missilä, A., Montemagni, S., Petrov, S., Pyysalo, S., Silveira, N., Simi, M., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). *Universal Dependencies 1.0*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, Universal Dependencies, 1.0. *Handle*: hdl:11234/1-1464.

Schluter, N. and van Genabith, J. (2007). *Modified French Treebank*. National Centre for Language Technology. URL: `http://doras.dcu.ie/15265/`.

Xue, N., Jiang, Z., Zhong, X., Palmer, M., Xia, F., Chiou, F.-D., and Chang, M. (2010). *Chinese Treebank 7.0*. Linguistic Data Consortium, Chinese Treebank, 7.0. ISLRN 156-627-429-482-3.