Une note sur l'analyse du constituant pour le français

Jungyeul Park
CONJECTO, 74 rue de Paris, 35000 Rennes, France
http://www.conjecto.com

RÉSUMÉ

Cet article traite des analyses d'erreurs quantitatives et qualitatives sur les résultats de l'analyse syntaxique des constituants pour le français. Pour cela, nous étendons l'approche de Kummerfeld *et al.* (2012) pour français, et nous présentons les détails de l'analyse. Nous entraînons les systèmes d'analyse syntaxique statistiques et neuraux avec le corpus arboré pour français, et nous évaluons les résultats d'analyse. Le corpus arboré pour le français fournit des étiquettes syntagmatiques à grain fin, et les caractéristiques grammaticales du corpus affectent des erreurs d'analyse syntaxique.

ABSTRACT

A Note on constituent parsing for French.

This paper deals with the quantitative and qualitative error analysis on French constituent parsing results. To this end, we extend the approach of Kummerfeld *et al.* (2012) to the French treebank for parser error analysis, and present details of the analysis for French. We train statistical and neural parsing systems, and evaluate parsing results using the French treebank. The French treebank provides fine-grained phrase labels and grammatical characteristics of the French treebank affect parsing errors.

MOTS-CLÉS : Analyse du constituant, corpus arboré, erreurs d'analyse syntaxique, systèmes d'analyse syntaxique statistiques et neuraux, français.

KEYWORDS: Constituent parsing, treebank, parsing errors, statistical and neural parsing systems, French.

1 Constituent Parsing for French

Treebanks, collections of parsed and syntactically annotated corpora, constitute an essential resource for natural language processing in any given language. The automatic syntactic analysis of sentences directly benefits from syntactically annotated corpora. Currently, most of the state-of-the-art parsers use the statistical or neural parsing approaches. These parsers use annotated syntactic information in the treebank to train parsing models. Several annotated phrase-structured treebanks have been created for French such as the French treebank (Abeillé *et al.*, 2003) and the Sequoia corpus (Candito & Seddah, 2012). Table 1 summarizes previous work on constituent parsing for French. This paper is intended to present several factors on constituent parsing for French including parsing results and an error analysis. We train and evaluate the French treebank (Abeillé *et al.*, 2003) using the state-of-art parsing systems : the statistical Berkeley parser (Petrov *et al.*, 2006) and the neural Trance parser (Watanabe & Sumita, 2015) (§ 2). Then, we extend Kummerfeld *et al.* (2012)'s parser error analysis to French (§ 3). Finally, we conclude the paper with discussion and future perspectives (§ 4).

Seddah *et al.* (2009)	84,93	using the Berkeley parser
Candito & Crabbé (2009)	88.29	gold POS + morphological clustering using Brown clustering
Candito & Seddah (2010)	87.80	gold lemma/POS + morphological clustering
Sigogne *et al.* (2011)	85.22	integrating the Lexicon-Grammar
Le Roux *et al.* (2014)	83.80	recognizing MWEs usnig CRFs and dual decomposition
Durrett & Klein (2015)	81.25	neural CRF parsing for multilingual settings
Coavoux & Crabbé (2016)	80.56	transition-based parsing with dynamic oracle (order-0 head-markovization)
Cross & Huang (2016)	83.31	transition-based parsing with dynamic oracle (no binarization)

TABLE 1 – Brief description and results of previous work on constituent parsing for French : Le Roux *et al.* (2014), Durrett & Klein (2015), Coavoux & Crabbé (2016) and Cross & Huang (2016) are based on a corpus split proposed in Seddah *et al.* (2013).

The main contribution of this paper is as follows. First, we explore various settings to parse the French treebank including parsing with functional information. Secondly, we propose parsing errors analysis for French based on Kummerfeld *et al.* (2012) to present the quantitative and qualitative error analysis. The error analysis script for French is publicly available at `https://github.com/jungyeul/taln2018`.

2 Experiments and Results

The current available version of the French treebank contains 45 files and 21,550 sentences (Abeillé *et al.*, 2003). [1] We use a corpus split proposed in Seddah *et al.* (2013) for training, development and test datasets directly from the French treebank instead of the distribution version from the SPMRL 2013 Shared Task. [2] This is mainly to train/evaluate the treebank using the different annotation such as training with functional information. While there are more sentences in the current treebank with 17,774/1,235/2,541 sentences for training/dev/evaluation, we use the exact data split from (Seddah *et al.*, 2013) (14,759/1,235/2,541). For statistical parsing using the Berkeley parser (Petrov *et al.*, 2006) [3], we report evaluation results using grammars which give the best results on development data. While the original Berkeley parser proposed several runs of training because of the EM algorithm which can find locally maximum likelihood parameters, we empirically found that each run of training gives same results. Therefore, we use the single run of training using the Berkeley parser with the default option. For experiments in this paper, we use Penn treebank-like preprocessing, especially by removing null elements (`*T*`) and functional information in the phrase label (*e.g.* `-SUJ` or `-OBJ`) as described in (Bikel, 2004). We evaluate the parser accuracy with the standard F_1 metric from `EVALB`. [4] While the SPMRL shared task provides the alternative `EVALB` [5], it produces the same F_1 scores for French. We only change the original `evalb` to display results for sentences $\leq$70 as in the

1. `http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php`
2. `http://www.spmrl.org/spmrl2013-sharedtask.html`
3. `https://github.com/slavpetrov/berkeleyparser`
4. `http://nlp.cs.nyu.edu/evalb`
5. `http://pauillac.inria.fr/~seddah/evalb_spmrl2013.tar.gz`

	berkeley+r	berkeley+f
(w/o gold POS)	79.26 (81.51)	77.02 (79.59)
(w/ gold POS)	80.95 (83.37)	78.55 (81.25)
# of NT label type	23	111

TABLE 2 – Parsing results using the statistical parser and the number of phrase non-terminal label types. For parsing results we also present F1 scores for sentences $\leq$70 in parentheses.

	trance+r	trance+f
(w/o gold POS)	78.05 (80.77)	76.39 (79.03)

TABLE 3 – Parsing results using the neural parser

shared task. We rename phrase labels which share the same label names with POS labels (usually for multi-word expressions or compound words) (+r). For example, we convert [$_{\mathrm{P}}$ [$_{\mathrm{P}}$ *D'*][$_{\mathrm{P}}$ *après*]] into [$_{\mathrm{P+}}$ [$_{\mathrm{P}}$ *D'*][$_{\mathrm{P}}$ *après*]] to differentiate between `P`s in the phrase label and the POS label. Therefore, we rename POS labels `A`, `ADV`, `C`, `CL`, `D`, `ET`, `I`, `N`, `P`, `PRO`, and `V` which also appear in the phrase labels. We note that the treebank of the SPMRL shared task has a similar annotation for compound words. For comparison reason, we also use functional information during training (+f) without renaming phrase labels. For example `np+suj` and `vppart+mod` instead of `np` and `vppart` are used for (+f). Table 2 shows the current parsing results on evaluation data by the Berkeley parser. Table 2 also shows the number of non-terminal (NT) label type without considering POS labels, in which `berkeley+r` has 12 phrase labels and 11 POS labels (renamed with +). We convert proposed alternative treebank forms (+r and +f) into the original preprocessed form without renaming and functional information to evaluate the result. We present the final scores from evaluation data based on best parsing results of dev data.

For neural parsing, we use the Trance parser (Watanabe & Sumita, 2015) [6] and a pre-trained 300 dimension embedding vector provided by Bojanowski *et al.* (2017) [7]. We use default options with 50 epochs for the Trance parser. Table 2 shows the current parsing results on evaluation data by the Trance parser.

3 Parsing Error Analysis

Recent state of the art parsing techniques are easily trained and evaluated if the syntactically annotated treebank is available. Their results, however, can be difficult to understand because grammars are automatically induced from the treebank. Kummerfeld *et al.* (2012) presented an approach to quantify constituent parsing errors based on the treebank annotation. [8] In this section, we extend Kummerfeld's approach to the French treebank parsing for parser error analysis. Error analysis is based on parsing results (+r). Table 4 shows the quantified number of each error w/o gold POS and w/ gold POS for the Berkeley and the Trance parsers.

6. `https://github.com/tarowatanabe/trance`
7. `https://fasttext.cc/docs/en/pretrained-vectors.html`
8. `https://github.com/jkkummerfeld/berkeley-parser-analyser`

	PP	NP	VP	MD	CL	PR	CO	SW	DL	UN	NI	UD
w/o (B)	2,036	531	380	195	301	17	479	1,885	681	678	444	3,302
w/ (B)	1,953	562	381	171	294	9	673	1,459	435	640	411	3,052
w/o (T)	1,956	593	338	310	282	16	436	1,765	617	728	559	3,841

TABLE 4 – Quantitative error analysis for the Berkeley parser :(B) for the Berkeley parser and (T) for the Trance parser with (w/) and without (w/o) gold POS labels. MD for modifier, CL for clause, PR for pronoun, CO for co-ordination, SW for single word, DL for different label, UN for unary, NI for np internal, and UD for undefined errors.

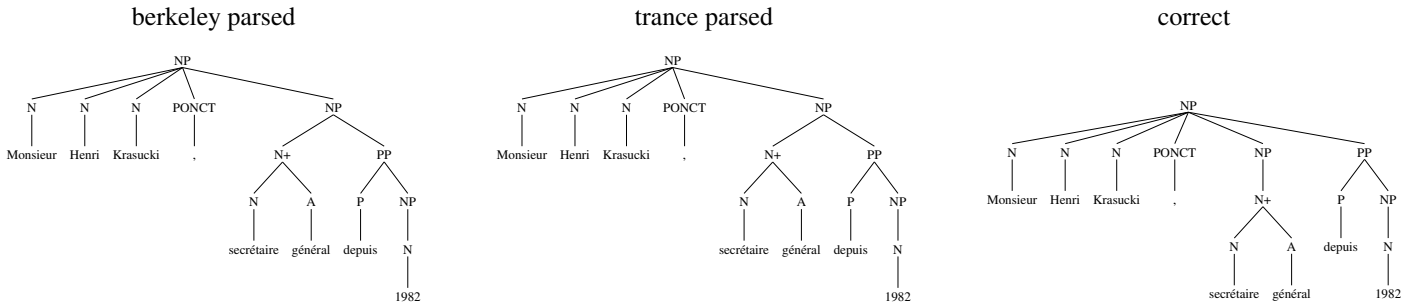

FIGURE 1 – PP attachment error : since `pp` is wrongly recognized as an argument of the sister `np` node instead an argument of its parent in (1), `pp` is low.

Attachment errors Attachment errors are the most frequent errors in constituent parsing for French (over 36% of parsing errors). They generally consist of mistakes and inconsistencies for recognizing arguments of the lexical head. There are six types of attachment errors : `pp`, `np`, `vp` (for `vn`, `vpinf` and `vppart`), `modifier` (for `ap` and `adp`), `clause` (for `sint`, `srel` and `ssub`), and `pron` (for `cl` and `pro`). See Figure 1 for an example of the PP attachement error.

(1) a. * [NP [N M.] [N Henri] [N Krasucki] [PONCT ,] [NP [N+ [N secrétaire] [A général]] [PP [P depuis] [NP [N 1982]]]]]

b. [NP [N M.] [N Henri] [N Krasucki] [PONCT ,] [NP [N+ [N secrétaire] [A général]]] [PP [P depuis] [NP [N 1982]]]]

Co-ordination error Annotating phrase with co-ordination in French is a difficult problem (*inter alia* Mouret (2007)). The current annotation in the French treebank shows a hierarchical structure, which is different with the English Penn treebank (a flat structure). Finding the correct scope of the coordinating conjunction is challenging, and co-ordination errors occur frequently. See Figure 2 for an example of the co-ordination error.

(2) a. * [PP [P d'] [NP [N ordre] [AP [A économique] [COORD [C et] [AP [A financier]]]]]]

b. * [PP [P d'] [NP [N ordre] [AP [A économique]]]] [COORD [C et] [AP [A financier]]]

c. [PP [P d'] [NP [N [N ordre] [A économique]] [COORD [C et] [AP [A financier]]]]]

Different label A phrase label is wrongly assigned. We note that POS label errors are not counted, and even for parsing with gold POS label, the Berkeley parser does not always obtain 100% for POS labeling accuracy. See Figure 3 for an example of the different label error.

(3) a. * [PP ... [NP [N sommes] [ADV+ [P en] [N jeu]]]]

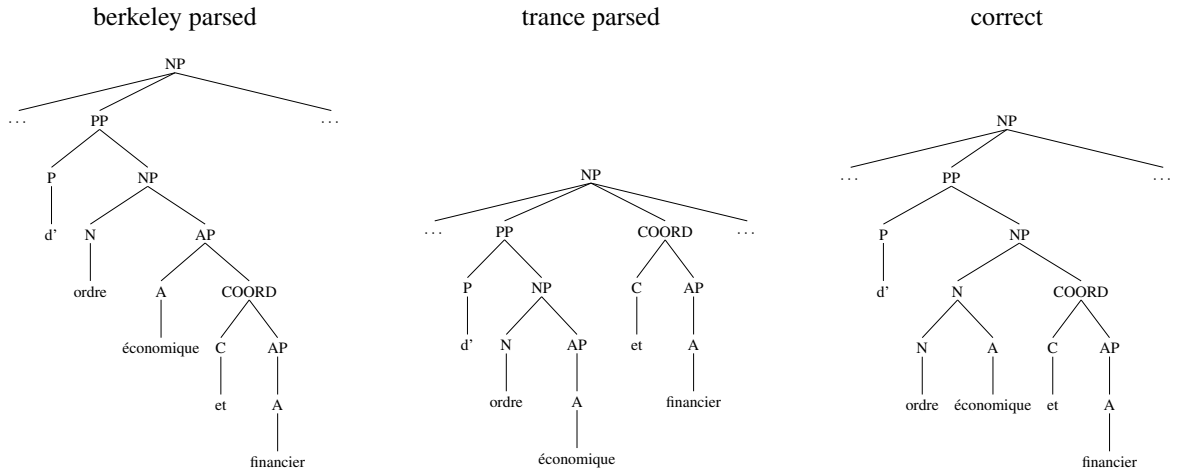

FIGURE 2 – Co-ordination error : `coord` is either low (B) or high (T). A coordinator *et* links with *économique* (B) or *d'ordre économique* (T) in (2).

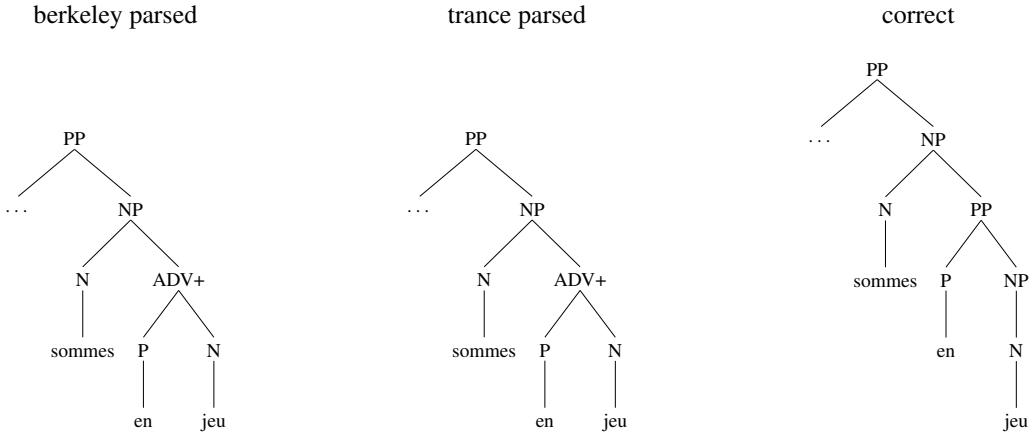

FIGURE 3 – Different label : `adv+` is wrongly recognized for `pp` in (3). It implies another error in which `n` for *jeu* is high (unary error).

b. [PP ... [NP [N sommes] [PP [P en] [NP [N jeu]]]]]

NP internal structure A general structure of the French treebank is relatively flat for the inside of NP as well as the entire sentence. For example, a sentence in (4) is an NP with a flat structure as follows : [NP [D ...] [N+ ...] [AP ...] [PP ...]]. However, both parsers fail to capture the flat structure for NP including a phrase segmentation. See Figure 4 for an example of the NP internal structure error.

(4) a. * [NP [D son] [N droit] [PP [P de] [NP [N préemption] [AP [A possible]]]] [PP [P sur] [NP [D le] [A futur] [N canal] [VPPART [V libéré]]]]]

b. * [NP [D son] [N droit] [PP [P de] [NP [N préemption] [AP [A possible]] [PP [P sur] [NP [D le] [A futur] [N+ [N canal] [A libéré]]]]]]]

c. [NP [D son] [N+ [N droit] [P de] [N préemption]] [AP [A possible]] [PP [P sur] [NP [D le] [A futur] [N canal] [VPPART [V libéré]]]]]

We do not detail single word and unary errors because they are mostly parts of another errors. Over 30% of parsing errors are undefined. We need to investigate these other error types for constituent parsing results, which can be more pertinent for French. We leave this for future work.

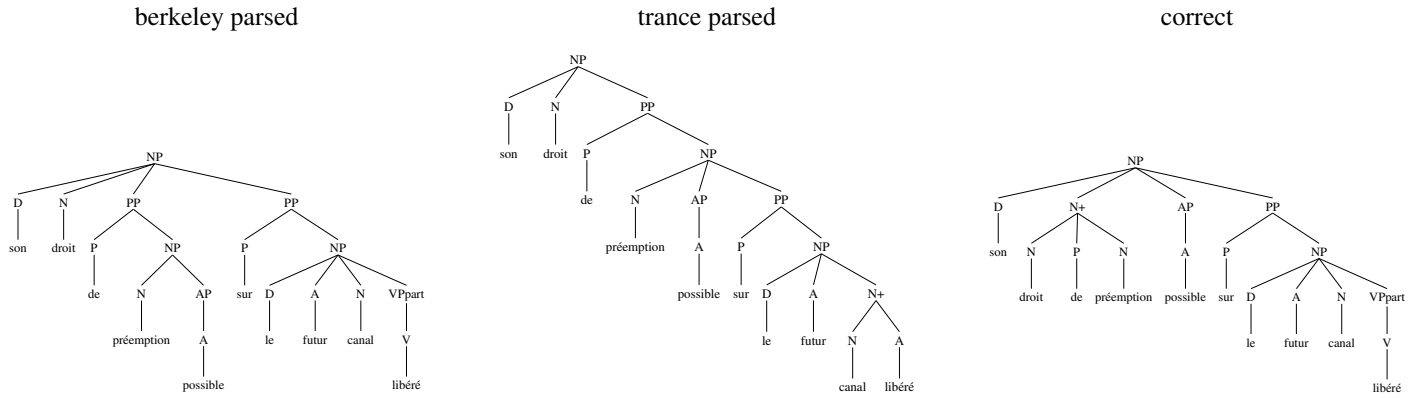

FIGURE 4 – NP internal error : `np` is wrongly constructed (4). It also implies another errors such as MWE recognition for *droit de préemption*, and `ap` (*possible*) and `pp` (*sur le futur ...*) attachment errors.

Previously, Sagot & de la Clergerie (2006) proposed an error mining technique based on parsing results from the FRMG (Thomasset & De la Clergerie, 2005) and the SXLFG (Boullier & Sagot, 2005) parsers. Since they parsed the raw corpus without knowing the correct parsed tree, they tried to find "suspicious" parsed results. These suspicious parsing trees are calculated based on predefined syntactic-related resources such as a morphological and syntactic lexicon L*efff* (Sagot, 2010) and a pre-syntactic processing pipeline SXPIPE (Sagot & Boullier, 2005).

4 Discussion and Conclusion

This paper dealt with error analysis studies on French constituent parsing results. While a neural parser improved parsing results for other languages such as English and Chinese, we did not obtain the better results for French. There are many intrinsic (learning rate, dropout, # of epochs, etc.) and extrinsic (word embedding and its dimension size) factors. Since training using 50 epochs takes over three or four days to learn a parsing model on a single machine, it wouldn't be easy to find proper parameters for French for neural parsing. We leave finding optimal parameter for French to future work. Functional information would also improve parsing results for certain morphologically rich languages (Chung *et al.*, 2010). The French treebank provides fine-grained phrase labels (111 different labels) and the Berkeley parser also generates additional internal phrase labels during training. PCFG rules in `berkeley+f` contain over 2M, compared to 0.4M in `berkeley+r` (*cf.* 18K vs. 15K for `trance+f` and `trance+r`). Such diversities with phrase labels can give a biased distribution. Therefore, functional information hardly effects or even tends to worsen parsing results in many cases. Investigating the effective way on clustering phrase labels can be one direction to improve parsing results and we leave this for future work. Instead of renaming phrase labels with `+`, we can also consider renaming with existing `*p`-like labels such as `np` or `pp` : *e.g.* the phrase label of compound words in [$_{\text{NP}}$ [$_{\text{N}}$ [$_{\text{N}}$ *banques*] [$_{\text{A}}$ *centrales*]]] is converted into [$_{\text{NP}}$ [$_{\text{NP}}$ [$_{\text{N}}$ ...] [$_{\text{A}}$ ...]]] instead of [$_{\text{NP}}$ [$_{\text{N+}}$ [$_{\text{N}}$ ...] [$_{\text{A}}$ ...]]]. Using `*p`-like renaming labels (12 phrase labels), we can find additional repetitive unary branches and we remove them during preprocessing. Converting into `*p`-like labels is straightforward except for `D` in which it would be `np` for numbers such as [$_{\text{NP}}$ [$_{\text{D}}$ *vingt*] [$_{\text{PONCT}}$ -] [$_{\text{D}}$ *cinq*] [$_{\text{D}}$ *mille*]] ; otherwise, `pp`. Consequently, we have 284,107 non-terminal nodes instead of 288,374 as in `berkeley+r` (excluding pre-terminal POS labels) in training data, and obtain only up to 75.42% F_1 score. This reflects the fact that recognizing multi-word expressions

(MWEs) and compound words is important in parsing for French, and it already proved in Le Roux *et al.* (2014) where they employed external linguistic resources such as DELAC, compound word dictionary for French (Courtois *et al.*, 1997) [9]. Exploring MWEs can be another direction to improve parsing results. [10] We note that we obtained slightly better results using the Berkeley parser than what the SPMRL shared task reported (gold setting) : 80.38 and 81.76 for w/o and w/ gold POS labels. This is probably because a preprocessing step for treebank data could be "slightly" dissimilar. We used our own the preprocessed French treebank to explore the different treebank settings.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a Treebank for French. In *Treebanks*, p. 165–188. Kluwer.

BIKEL D. M. (2004). Intricacies of Collins' Parsing Model. *Computational Linguistics*, **30**(4), 479–511.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

BOULLIER P. & SAGOT B. (2005). Efficient and Robust LFG Parsing : SxLFG. In *Proceedings of the Ninth International Workshop on Parsing Technology (IWPT2005)*, p. 1–10, Vancouver, British Columbia : Association for Computational Linguistics.

CANDITO M. & CRABBÉ B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, p. 138–141, Paris, France : Association for Computational Linguistics.

CANDITO M. & SEDDAH D. (2010). Parsing Word Clusters. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 76–84, Los Angeles, CA, USA : Association for Computational Linguistics.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 321–334, Grenoble, France : ATALA/AFCP.

CHUNG T., POST M. & GILDEA D. (2010). Factors Affecting the Accuracy of Korean Parsing. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 49–57, Los Angeles, CA, USA : Association for Computational Linguistics.

COAVOUX M. & CRABBÉ B. (2016). Neural Greedy Constituent Parsing with Dynamic Oracles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 172–182, Berlin, Germany : Association for Computational Linguistics.

COURTOIS B., GARRIGUES M., GROSS G., GROSS M., JUNG R., MATHIEU-COLAS M., MONCEAUX A., ANNE P.-M., SILBERZTEIN M. & VIVÈS R. (1997). *Dictionnaire électronique DELAC : les noms composés binaires*. Rapport interne, Université Paris 7, Paris.

CROSS J. & HUANG L. (2016). Span-Based Constituency Parsing with a Structure-Label System and Provably Optimal Dynamic Oracles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 1–11, Austin, Texas : Association for Computational Linguistics.

9. http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/delac.html
10. https://typo.uni-konstanz.de/parseme

DURRETT G. & KLEIN D. (2015). Neural CRF Parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 302–312, Beijing, China : Association for Computational Linguistics.

KUMMERFELD J. K., HALL D., CURRAN J. R. & KLEIN D. (2012). Parser Showdown at the Wall Street Corral : An Empirical Investigation of Error Types in Parser Output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 1048–1059, Jeju Island, Korea : Association for Computational Linguistics.

LE ROUX J., ROZENKNOP A. & CONSTANT M. (2014). Syntactic Parsing and Compound Recognition via Dual Decomposition : Application to French. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 1875–1885, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.

MOURET F. (2007). *Grammaire des constructions coordonnées. Coordinations simples et coordinations à redoublement en français contemporain.* PhD thesis, Université Paris 7 - Denis Diderot.

PETROV S., BARRETT L., THIBAUX R. & KLEIN D. (2006). Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, p. 433–440, Sydney, Australia : Association for Computational Linguistics.

SAGOT B. (2010). The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

SAGOT B. & BOULLIER P. (2005). From raw corpus to word lattices : robust pre-parsing processing. In *Proceedings of the 2nd International Conference Language And Technology (L&T'05)*, Poznań, Pologne.

SAGOT B. & DE LA CLERGERIE E. V. (2006). Error Mining in Parsing Results. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, p. 329–336, Sydney, Australia : Association for Computational Linguistics.

SEDDAH D., CANDITO M. & CRABBÉ B. (2009). Cross parser evaluation : a French Treebanks study. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, p. 150–161, Paris, France : Association for Computational Linguistics.

SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J. D., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA GALLETEBEITIA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIÓRKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & DE LA CLERGERIE E. V. (2013). Overview of the SPMRL 2013 Shared Task : A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.

SIGOGNE A., CONSTANT M. & LAPORTE E. (2011). French parsing enhanced with a word clustering method based on a syntactic lexicon. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, p. 22–27, Dublin, Ireland : Association for Computational Linguistics.

THOMASSET F. & DE LA CLERGERIE E. V. (2005). Comment obtenir plus des métagrammaires. In *Proceedings of the Conference TALN 2005*, Dourdan, France : ATALA/AFCP.

WATANABE T. & SUMITA E. (2015). Transition-based Neural Constituent Parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1169–1179, Beijing, China : Association for Computational Linguistics.