Ubiquitous Usage of a French Large Corpus: Processing the Est Republicain Corpus

Djamé Seddah[1,2], **Marie Candito**[1], **Benoit Crabbé**[1] **& Enrique Henestroza Anguiano**[1]

1. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, 175 rue du Chevaleret, 75013 Paris, France
2. Université Paris Sorbonne, 7 rue Victor Cousin, 75006, Paris
firstname.lastname@inria.fr

Abstract

In this paper, we introduce a set of resources that we have derived from the EST RÉPUBLICAIN CORPUS, a large, freely-available collection of regional newspaper articles in French, totaling 150 million words. Our resources are the result of a full NLP treatment of the EST RÉPUBLICAIN CORPUS: handling of multi-word expressions, lemmatization, part-of-speech tagging, and syntactic parsing. Processing of the corpus is carried out using statistical machine-learning approaches - joint model of data driven lemmatization and part-of-speech tagging, PCFG-LA and dependency based models for parsing - that have been shown to achieve state-of-the-art performance when evaluated on the French Treebank. Our derived resources are made freely available, and released according to the original Creative Common license for the EST RÉPUBLICAIN CORPUS. We additionally provide an overview of the use of these resources in various applications, in particular the use of generated word clusters from the corpus to alleviate lexical data sparseness for statistical parsing.
Keywords: Corpus, parsing, out-of-domain

1. Introduction

Most supervised methods currently in use in Natural Language Processing (NLP) are built upon annotated corpora. Such resources are extremely costly to develop, even with the growing use of crowd-sourced development methods, which are still to be proven useful for building complex annotations. However, a compromise can be achieved by using high quality automatic annotation applied to large corpora in order to provide data for semi-supervised methods. Among many other uses, such corpora have been successfully used to increase statistical parsing performance in a self-training setting (McClosky et al., 2006), to acquire word clusters for use as features in dependency parsing (Koo et al., 2008), or for lexical acquisition (Chrupała and van Genabith, 2007)

Specifically, since the release of the EST RÉPUBLICAIN CORPUS, a freely-available large collection of regional news press articles of French, compiled and released by the CNRTL[1][2] (Gaiffe and Nehbi, 2009), a new range of work based on the use of unsupervised and semi-supervised learning techniques has emerged, and has aided in significantly improving the state-of-the-art of statistical parsing for French (Candito and Crabbé, 2009; Candito and Seddah, 2010; Candito et al., 2011; Le Roux et al., 2011).

Outside of NLP, a growing field of interest for this type of data is Experimental Linguistics, especially when applied to resource-poor languages such as French. For example, an automatically lemmatised and POS-tagged form of the EST RÉPUBLICAIN CORPUS corpus has recently been used for an empirical linguistic study by Thuilier et al. (2010), where the author extracts noun-adjective association scores (chi square value, fisher association scores and frequency counts) according to their relative positions. This information is then used to model preferential choices in the positioning of the attributive adjective in French, with the goal of identifying relevant factors made by French speakers in this situation.

In this paper, we introduce the set of resources we derived from the EST RÉPUBLICAIN CORPUS, with a focus on their use in creating state-of-the-art statistical parsers for French. These resources are freely available, and released according to the original Creative Common license for the EST RÉPUBLICAIN CORPUS.

2. Introducing the Est Républicain Corpus

The EST RÉPUBLICAIN CORPUS (henceforth ERC) is a local newspaper based in the eastern part of France, covering the Lorraine and Franche Conté geographic region. It contains around 149 million words, and more than 9.2 million sentences. Collecting only local news and important events of interests, this corpus cannot be considered as being balanced, unlike the British National Corpus (Leech, 1992) or the American National Corpus (Ide and Macleod, 2001), for instance.

Having been built with experimental linguistic usage in mind by the CNRTL, the ERC strictly follows the Text Encoding Intiative standards (TEOI, (Sperberg-McQueen and Burnard, 1994)), its structure being described at length by Nehbi and Gaiffe (2009).[3]

3. Processing the Est Republicain

Our primary use for this corpus is to use it as a source of data for the reduction of lexical sparseness issues, which are inherent to the statistical parsing of small sized treebanks such as the French Treebank (FTB, (Abeillé et al., 2003)).

3.1. Data set

Before presenting the various treatments we applied on the ERC, we introduce briefly the FTB. THE FRENCH TREEBANK is the first annotated and manually corrected treebank for French. The data is annotated with labeled con-

[1] Centre National de Ressources Textuelles et Lexicales

[2] http://www.cnrtl.fr/corpus/estrepublicain/

[3] http://www.cnrtl.fr/corpus/estrepublicain/est-documentation.php

stituent trees augmented with morphological annotations and functional annotations of verbal dependents. Its key properties, compared with the PTB, are the following :

Size: The FTB consists of 350,931 tokens and 12,351 sentences, that is less than a third of the size of PTB. The average length of a sentence is 28.41 tokens. By contrast, the average sentence length in the Wall Street Journal section of the PTB is 25.4 tokens.

A Flat Annotation Scheme: Both the FTB and the PTB are annotated with constituent trees. However, the annotation scheme is flatter in the FTB. For instance, there are no VPs for finite verbs and only one sentential level for clauses or sentences whether or not they are introduced by a complementizer. Only the *verbal nucleus* (VN) is annotated and comprises the verb, its clitics, auxiliaries, adverbs and negation.

Inflection: French morphology is richer than English and leads to increased data sparseness for statistical parsing. There are 24,098 lexical types in the FTB, with an average of 16 tokens occurring for each type.

Compounds: Compounds are explicitly annotated and very frequent in the treebank: 14.52% of tokens are part of a compound. Following Candito and Crabbé (2009), we use a variation of the treebank where compounds with regular syntactic patterns have been expanded. We refer to this instance as FTB-UC.
They include digit numbers (written with spaces in French) (e.g. *10 000*), frozen compounds (eg. *pomme de terre 'potato'*) but also named entities or sequences whose meaning is compositional but where insertion is rare or difficult (e.g. *garde d'enfant 'child care'*).
As noted by Arun and Keller (2005), compounds in French may exhibit ungrammatical sequences of tags as in *à la va vite* 'in a hurry': Prep+ Det+ finite verb + adverb or can include "words" which do not exist outside a compound (e.g. *hui* in *aujourd'hui* 'today'). Therefore, Compounds receive a two-level annotation: constituent parts are described at an embedded level using the same POS tag set as the compound POS.

FTB-UC "CC" **tagset:** This is the part-of-speech tagset developed by (Crabbé and Candito, 2008) (Table 1), known to provide the best constituency parsing performance for French (Seddah et al., 2009). Like in the FTB, preterminals are the main categories, but they are also augmented with a WH flag for A, ADV, PRO and with the mood for verbs (there are 6 moods). No information is propagated to non-terminal symbols.

ADJ ADJWH ADV ADVWH CC CLO CLR CLS CS DET DETWH ET I NC NPP P P+D P+PRO PONCT PREF PRO PROREL PROWH V VIMP VINF VPP VPR VS

Table 1: CC tag set

3.2. Preprocessing and Tokenization

As the FTB is our primary source of annotated data and the main beneficiary of any improvement coming from the ERC, it is important that the ERC and the FTB share the same tokenization with respect to punctuation marks and word forms. Punctuation matters, as was indeed shown by Foster et al. (2007): converting the BNC punctuation set (Leech et al., 1994) to the Penn Treebank style was shown to boost actual parsing performance on out-of-domain text.

Regarding word forms, it should be noted that unlike many widely used treebanks, the FTB contains a large amount of multi-word expressions (or word compounds, to use the FTB terminology). Those expressions range from complex prepositions such as *au sein de* (within) or *alors même que* (even though) to named entities, *Banque Européenne de Reconstruction et de Développement* (European Bank for Reconstruction and Development). In order to match this particularity of the FTB, we applied a word compound recognition process to the ERC using compounds extracted from a specific version FTB-UC of the FTB, where compounds with regular syntax have been undone, leaving only the 250 most frequents ones in the treebank.[4] This phase is achieved using the pre-processing tools part of the BONSAÏ package (Candito and Crabbé, 2009).

3.3. POS Tagging and Data-Driven Lemmatisation

The first step toward obtaining high quality part-of-speech tag annotation is to use models that can handle the French language's rich morphology while providing state-of-the-art performance.
In order to assign morphological tags and lemmas to words, we use a variation of the MORFETTE model described in (Chrupała et al., 2008) and adapted to French by Seddah et al. (2010) It is a sequence-labeling model which combines the predictions of two classification models (one for morphological tagging and one for lemmatization) at decoding time, using a beam search. While (Chrupała et al., 2008) use Maximum Entropy training to learn P_M and P_L, we use the MORFETTE models described in (Seddah et al., 2010), that are trained using the Averaged Sequence Perceptron algorithm (Freund and Schapire, 1999).The two classification models incorporate additional features calculated using the Le*fff* lexicon (Sagot, 2010).
Table 2 shows detailed results on the development and test sets of the FTB-UC [5], when MORFETTE is trained on the FTB-UC training set. To the best of our knowledge, the part-of-speech tagging performance reported here is state-of-the-art for Frenchand the lemmatization performance has no comparable results.[6]
To evaluate the accuracy of this process on the ERC, 433 sentences have been randomly sampled from the corpus and jointly lemmatized and POS-tagged by MORFETTE. Subsequently, a manual validation step was carried out by a pair of annotators. As shown in Table 3, MORFETTE's POS tagging accuracy on the ERC is slightly inferior than on its original data set. This demonstrates that even though the ERC does not originate from the same source as the FTB,

[4]The list of compounds is available at: `http://alpage.inria.fr/statgram/frdep/mwes.txt`.

[5]First 10% for the test set, next 10% for the development set and the rest for training.

[6]Please note that this evaluation was performed on a subset of the ERC gold standard (Table 4) with hand annotated lemmas.

Dev set	Overall	Unseen (4.8%)
POS acc	97.38	91.95
Lemma acc	98.20	92.52
Joint acc	96.35	87.16
Test set	Overall	Unseen (4.62 %)
POS acc	97.68	90.52
Lemma acc	98.36	91.54
Joint acc	96.74	85.28

Table 2: MORFETTE performance on the FTB-UC canonical development and test sets (with and without punctuation)

	Overall	Unseen (12.47%)
Pos acc	96.66	89.19
Lemma acc	98.37	93.74
Joint acc	96.07	86.24

Table 3: MORFETTE performance on the ERC gold standard set

	EST RÉP.	FTB	
	GOLD STD	DEV	TRAIN
# Sentences	529	1235	9881
Avrg. Length	21	29.6	28.1
Std. deviation	12.9	16	16.5
Counts using any token type (including punct.)			
Vocabulary size	3337	7222	24110
% of unknown	29.2	22.5	-
# occurrences	11114	36508	278k
% of unknown	11.2	5.2	-
% of proper nouns	5.1	4.1	4.0
Counts using lower-cased alphanumerical tokens			
Vocabulary size	3173	6904	22526
% of unknown	28	21.06	-
# occurrences	9552	30940	235k
% of unknown	12.1	5.7	-

Table 4: Properties of the ERC gold standard

which is derived from the LE MONDE newspaper, there is apparently little lexical variation between those two corpus. Alternatively, one could also argue that any variation between the two sources is easily captured by the MORFETTE joint model of lemmatization and POS tagging.

4. Statistical Parsing Evaluation

To evaluate parsing performance on the ERC, we used a slightly extended version of the test set presented above (which includes around 100 sentences more). This corpus is part of a set of out-of-domain corpora produced by the SEQUOIA[7], see (Candito and Seddah, 2012) for details. Table 4 summarises the main properties of the ERC gold standard. For the purposes of comparison, the FTB's main characteristics are displayed in the rightmost column.

We applied two different syntactic parsing aproaches in order to parse the ERC: (i) a constituency parsing PCFG-LA based tool chain (BONSAÏ, (Candito et al., 2010a)) and (ii) a transition based dependency parsing chain (MALT, (Nivre et al., 2007)).

4.1. Constituency parsing: a PCFG-LA architecture

For PCFG-LA parsing, we follow the word clustering strategy described in (Candito and Crabbé, 2009; Candito and Seddah, 2010) that consists in first generating unsupervised word clusters (Brown et al., 1992) using the Liang (2005) implementation, then training a PCFG-LA parser (Petrov et al., 2006) on a treebank where word forms have been replaced by word clusters. Those are then replaced in the parsed data by the original tokens before being subjected to a functional labeling step (Candito et al., 2010b). This method provides state-of-the-art results (Candito et al., 2010a; Le Roux et al., 2011).

Different methods of morphological clustering are possible (clustering built on pure word forms, lemmas, disinflected word forms, etc.) and achieve a similar range of performance results (Candito and Seddah, 2010). The interest in using clusters built on morphologically-processed corpora is that this approach alleviates data sparseness issues. In fact, when a parser faces clustered data, the notion of an unknown word becomes almost meaningless. In the FTB-UC test set, less than 0.6% of tokens are unknown after clustering. This can explain the very high performance of PCFG-LA parsing when trained on word clusters: the lexicon is drastically reduced (only 1,700 clustered tokens in the FTB versus 7,052 other words). Results of our PCFG-LA based are shown in Table 5. In addition to the classical PARSEVAL metrics (Black et al., 1991), we also provide leaf-Ancestor accuracy measure (Sampson and Babarczy, 2003).

Given the small size of the test set, definitive conclusions are difficult to draw, nevertheless performance seem to be in par with previously reported results on the FTB. On a similar setting (PCFG-LA and unsupervised word clustering), Candito and Seddah (2010) report a F-score of 88.22 % and 96.98 of POS accuracy on sentence of length <=40. Our own results, 86.82 % (POS: 94.92), confirm that the lexical gap between the ERC and the FTB is easily circumvented by our word clustering approach.

Sent. size	LR	LP	F_1	Leaf.	Pos Acc.
(<=40)	87.00	86.65	86.82	0.94	94.92
(all)	85.54	85.44	85.49	0.93	94.76

Table 5: Constituency evaluation of the BKY parser on the ERC gold standard

4.2. Dependency parsing: transition and graph based architecture

We processed the L'Est Republicain corpus using the BONSAI package. [8] It first performs segmentation and identification of multi-word-expressions, followed by part-of-speech tagging using the MElt tagger (Denis and Sagot,

[7] French ANR project ANR-08-EMER-013, 2009-2011

[8] alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

	LAS	UAS	LaS
BKY	84.36	88.07	89.54
Malt	82.68	86.76	88.98
MST	86.05	89.3	91.11

Table 6: Dependency evaluation of the ERC gold standard

2009), morphological analysis using the Lefff lexicon (Sagot, 2010), and parsing using MaltParser (Nivre et al., 2007), a linear-time transition-based dependency parser that is well suited for efficiently parsing large corpora. Results are shown Table 6. Those were calculated using the CONLL 2007 evaluation software. In addition to the Malt parser's results, we also include the evaluation of our constituency to dependency chain (using the BKY parser own tagging) and MsT results (Mcdonald et al., 2006). As in the previous approach we presented, the performance levels exhibited in previously reported results on French in domain dependency parsing (Candito et al., 2010a) and the one presented here are sufficiently close so we can assume a strong similarity between those domains. It shall be noted that MsT seems to be less sensitive to the small domain variation between the FTB and the ERC (LAS:86.05 here vs LAS:87.3). We leave for future work further investigations on that matter.

Parsed versions of the ERC using those parsers are freely available.

5. Conclusion

Having access to the EST RÉPUBLICAIN CORPUS has allowed us to improve our initial statistical parsing result by granting us access to a large amount of data that can alleviate data sparseness issues. The ERC, when morphologically clustered, can also act as an efficient bridge corpus between different domains. In recent work (Candito et al., 2011), we were able to bridge the lexical gap between the journalistic and biomedical text domains simply by adding comparatively little out-of-domain data (5 million words) to the initial ERC clustered data set.

Besides improving the state-of-the-art of French statistical parsing, we are particularly proud that an early release of a lemmatized version of this corpus has paved the way for research on experimental linguistics of French in our lab and university (Thuilier et al., 2010). The ERC has the potential to be very useful for research in NLP and other areas of linguistic study, and thus we are glad to make the set of resources we have produced (lemmatized, tagged, parsed, and clustered versions of the ERC) freely available to the research community [9].

6. References

Anne Abeillé, Lionel Clément, and François Toussenel, 2003. *Building a Treebank for French*. Kluwer, Dordrecht.

Abhishek Arun and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of french. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 306–313, Ann Arbor, MI.

E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 306–311, San Mateo (CA). Morgan Kaufman.

Peter F. Brown, Vincent J. Della, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 138–141, Paris, France, October. Association for Computational Linguistics.

M. Candito and D. Seddah. 2010. Parsing word clusters. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 76–84. Association for Computational Linguistics.

Marie Candito and Djamé Seddah. 2012. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de la 19ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, Grenoble, France, Juin.

M. Candito, J. Nivre, P. Denis, and E.H. Anguiano. 2010a. Benchmarking of statistical dependency parsers for french. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 108–116. Association for Computational Linguistics.

Marie Candito, Benoit Crabbé, and Pascal Denis. 2010b. Statistical french dependency parsing : Treebank conversion and first results. In *Proceedings of LREC'2010*, Valletta, Malta.

Marie Candito, Enrique Henestroza Anguiano, and Djamé Seddah. 2011. A word clustering approach to domain adaptation: Effective parsing of biomedical texts. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 37–42, Dublin, Ireland, October. Association for Computational Linguistics.

Grzegorz Chrupała and Josef van Genabith. 2007. Using very large corpora to detect raising and control verbs. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of LFG07*, Stanford University. CSLI Publications.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with morfette. In *In Proceedings of LREC 2008*, Marrakech, Morocco. ELDA/ELRA.

Benoit Crabbé and Marie Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)*, pages 45–54, Avignon,

[9] http://alpage.inria.fr/estrepu/

France.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proc. of PACLIC*, Hong Kong, China.

J. Foster, J. Wagner, D. Seddah, and J. Van Genabith. 2007. Adapting wsj-trained parsers to the british national corpus using in-domain self-training. In *Proceedings of the Tenth IWPT*, pages 33–35.

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296.

Bertrand Gaiffe and Kamel Nehbi. 2009. Le corpus de l'Est Républicain.

N. Ide and C. Macleod. 2001. The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, pages 274–280. Citeseer.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08*, pages 595–603, Columbus, USA.

Joseph Le Roux, Benoît Favre, Seyed Abolghasem Mirroshandel, and Alexis Nasr. 2011. Modèles génératif et discriminant en analyse syntaxique : expériences sur le corpus arboré de paris 7. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier, France, June.

G. Leech, R. Garside, and M. Bryant. 1994. Claws4: the tagging of the british national corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628. Association for Computational Linguistics.

G. Leech. 1992. 100 million words of english: the british national corpus. *Language Research*, 28(1):1–13.

Percy Liang. 2005. Semi-supervised learning for natural language. In *MIT Master's thesis*, Cambridge, USA.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA, June. Association for Computational Linguistics.

Ryan Mcdonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proc. of CoNLL'06*, New York, USA.

Kamel Nehbi and Bertrand Gaiffe. 2009. Tei est républicain : Encodage du corpustei p5. March.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 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*, Sydney, Australia, July. Association for Computational Linguistics.

Benoît Sagot. 2010. The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for french. In *Proceedings of LREC'10*, Valetta, Malta.

Geoffrey Sampson and Anna Babarczy. 2003. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9(04):365–380.

Djamé Seddah, Marie Candito, and Benoit Crabbé. 2009. Cross parser evaluation and tagset variation: A French Treebank study. In *Proceedings of the 11th Internation Conference on Parsing Technologies (IWPT'09)*, pages 150–161, Paris, France, October. Association for Computational Linguistics.

Djamé Seddah, Grzegorz Chrupała, Ozlem Cetinoglu, Josef van Genabith, and Marie Candito. 2010. Lemmatization and statistical lexicalized parsing of morphologically-rich languages. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

C.M. Sperberg-McQueen and Lou Burnard. 1994. *TEI guidelines for electronic text encoding and interchange (P3). Chicago and Oxford,* `http://-etext.virginia.edu/TEI.html`.

Juliette Thuilier, Gwendoline Fox, and Benoît Crabbé. 2010. Approche quantitative en syntaxe : l'exemple de l'alternance de position de l'adjectif épithète en français. In Procedings of Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Canada, juillet.*