Cheating a Parser to Death:
Data-driven Cross-Treebank Annotation Transfer

Djamé Seddah[1,2]   **Eric de la Clergerie[2]**   **Benoît Sagot[2]**
Héctor Martínez Alonso [4]   **Marie Candito[3]**

[1]Université Paris-Sorbonne   [2]Inria   [3]Université Denis Diderot & CNRS   [4] Thomson Reuters Lab
djame.seddah@paris-sorbonne.fr, {eric.de_la_clergerie, benoît.sagot}@inria.fr,
hector.martinezalonso@thomsonreuters.com, marie.candito@linguist.univ-paris-diderot.fr

Abstract

We present an efficient and accurate method for transferring annotations between two different treebanks of the same language. This method led to the creation of a new instance of the French Treebank (Abeillé et al., 2003), which follows the Universal Dependency annotation scheme and which was proposed to the participants of the CoNLL 2017 Universal Dependency parsing shared task (Zeman et al., 2017). Strong results from an evaluation on our gold standard (94.75% of LAS, 99.40% UAS on the test set) demonstrate the quality of this new annotated data set and validate our approach.

Keywords: Treebanking, Universal Dependencies, Syntax, Automatic Correction, Cross-annotation Transfer

1. Introduction

After many decades of treebanking initiatives (Einarsson, 1976; Marcus et al., 1993), the interest in developing annotated corpora no longer needs to be justified. Although a distinction can be noted between treebanks created for linguistic purposes and those only conceived in a natural language processing perspective, it tends to fade away in the face of the ever growing machine learning addiction to new sources of labeled data. In fact, not only can any annotated corpus be used as a primary or secondary source of training data within more or less complex systems, but hand-crafted syntactic resources such as grammars and lexicons can be used as sources of features to guide data driven systems (Øvrelid et al., 2009; Villemonte De La Clergerie, 2014a). The crucial point here lies in the interopability of such heterogenous sources of information. Before the rise of the Universal Dependency initiative (Nivre et al., 2017) and its eponymous scheme, henceforth UD, which resulted in the release of 81 treebanks on more than 50 language, the situation was at best complicated. Nevertheless, the pre-UD multitude of annotation schemes allowed many to use stacking methodologies for predicting syntactic annotations of a certain type and following specific guidelines (e.g. UD dependencies) with the help of other types of annotations that follow different schemes, sometimes even of a different topological nature (Farkas and Bohnet, 2012; Björkelund et al., 2013; Ambati et al., 2013; Ribeyre et al., 2015). In most cases, taking into account such heterogenous syntactic information in the form of additional features does improve parsing accuracy.

Unsurprisingly, the performance gain is generally outstanding whenever such features are extracted from gold annotations. When the goal is to produce new reference annotated data, such an performance gain results in fewer post-annotation corrections. In case of converting one treebank to another annotation scheme, such gold information is of course readily available and has the potential to considerably ease this process.

In this paper, we describe such a conversion effort, for which we had to meet with another drastic constraint; in the context of the preparation of the CoNLL 2017 shared task on "Multilingual Parsing from Raw Text to Universal Dependencies" (Zeman et al., 2017), we had less than two weeks for converting the French Treebank (Abeillé et al., 2003, hereafter FTB) in its SPMRL[1] dependency version (Seddah et al., 2013) into a new one that complies with the UD guidelines.

Such an objective forced us to think of all possible techniques that could help producing a treebank that would follow the UD scheme with the best possible accuracy. Since we were to produce a new data set, the use of a data-driven process fed with gold features whenever possible was the only way out. The result of our conversion process, as measured on a silver standard in terms of labeled attachment accuracy (LAS), reaches around 98.50% on the Sequoia UD Treebank (Candito and Seddah, 2012; Nivre et al., 2017). Against a smaller and manually validated subset, we reach 94.75% of LAS and 99.42 for unlabeled attachment score. These scores are likely to reflect the high quality of our resulting data set.

In the remaining of this paper, we describe the methodology we used to build the UD version of the FTB, hereafter FTB-UD, and present our evaluation process and results. The FTB-UD is available under the same licence conditions as the original FTB.[2]

2. Method Overview

The basic idea is the following: we had access to a rule-based system for automatically converting another treebank, namely the French Sequoia Treebank (Candito and Seddah, 2012, hereafter SEQUOIA), into UD. After adapting the FTB's native tokenization scheme to UD, this conversion system was directly applied to the FTB. This resulted in many errors: 16% of the sentences contained one or more errors at one or more levels (POS, dependency, head), between 6 and 7% of tokens were flagged as Fail-

---

[1]Statistical Parsing of Morphologically-Rich Languages.
[2]https://github.com/
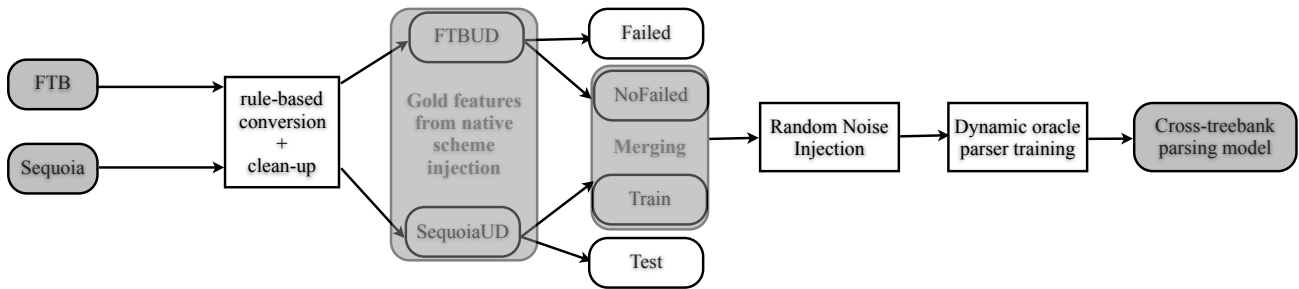UniversalDependencies/UD_French-FTB

Figure 1: Overview of our cross-treebank parser training process

ure conversion, leading of course to many more incorrect tree structures. The FTB being six times larger than SEQUOIA, adapting and extending the initial set of rules was not feasible in such a short time. We automatically corrected incorrect coordination tree structures and manually corrected missing POS resulting from conversion failures. We then decided to reparse all error-flagged dependencies using our robust shift-reduce parser with dynamic oracle (Villemonte De La Clergerie, 2013).

The idea was to build a pseudo gold training set (made of 90% of the SEQUOIA treebank and of the FTB training sentences that contained no conversion errors, leaving aside 20% of those for pseudo-gold evaluation) to which we injected both (i) external gold morpho-syntactic features coming from the FTB SPMRL version and (ii) random noise, such as empty dependencies, in the same proportions as the initial conversion errors (see Figure 1 for an overview of the training process). We then parsed all erroneous sentences (all incorrect edges were deleted) with this model with the hypothesis that the parser would be able to predict correct dependencies assuming the proper external gold features were to be provided.

3. Building the FTB-UD

Besides providing another source of annotated French data to the CoNLL 2017 shared task participants, our primary goal was to enable cross-parsing comparisons between different annotation schemes, namely the native FTB dependency scheme (Candito et al., 2010) as instantiated in the SPMRL shared tasks (Seddah et al., 2013; Seddah et al., 2014) and the then upcoming UD 2.0 scheme (Nivre et al., 2017) that was to be used for this shared task (Zeman et al., 2017). For these reasons, our starting point is the FTB SPMRL instance and not its latest incarnation.[3]

3.1. Multi-word Expression Treatment

We started by adapting the annotation scheme for multi-word expressions (MWEs). The treebank with less types of MWEs annotated is the Sequoia treebank, containing fixed functional MWEs. We thus used the existing rule-based software of Candito and Crabbé (2009) to "undo" non functional MWEs, namely to recover a regular syntactic structure for regular nominal, adjectival, verbal and adverbial MWEs. The patterns for spotting and undoing

MWEs are a subset of those of Candito and Crabbé (2009). All the remaining MWEs were then represented using the `fixed` dependency label, used for functional MWEs. This choice can be discussed in the light of the current debate within the UD community regarding the status to give to named entities. For example, the FTB contains many nominal named entities (tagged N N, e.g for persons), assuming a proper disambiguation step, those could have received a `flat:name` label instead. [4]

However, we then adapted the word segmentation to that of French UD 2.0, the main difference concerning amalgamated prepositions: e.g. the amalgamated preposition+determiner *au (litt. "to the")* is systematically treated as one token but two words (*à (to)* and *le (the)*).

3.2. Application of Sequoia to UD rule-based converter

Before we started this work, another research team was working on the conversion of the SEQUOIA treebank to the UD annotation scheme (Guillaume et al., to appear) using their graph rewriting engine (Guillaume et al., 2012). Because the SEQUOIA treebank native annotation scheme uses the same guidelines as the FTB, the use of the rule-set they developed was favored in order to bootstrap the conversion process. However, both corpora differ considerably in size (resp. 3k vs 18k sentences) and domains (wikipedia, europarl, biomedical for SEQUOIA, international and national news-wire for the FTB), leading the application of the SEQUOIA to UD conversion process to a new domain to be non-trivial. As we mentioned in the previous section, the resulting treebank contained 16% of sentences with one or more errors and 6% after correction of some coordinate structures. The next two sections describe how we corrected those errors.

3.3. POS Correction and Injection of Gold Features

POS-correction The application of the conversion rules resulted in a failure to produce a POS tag for 89 word-forms (61 in the training set, 3 in the development set, 25 in the test set). We manually reviewed and POS-annotated all these cases.

Injection of Morpho-syntactic Gold Features We first developed an algorithm for automatically post-align

---

[3]`http://ftb.linguist.univ-paris-diderot.fr`, released in December 2016.

[4]Please note that the version distributed for the UD Shared Task did not contain this regularization, which will be included in the next major release.

the output of the conversion with the original FTB SPMRL files, which differ in how they are segmented into tokens and wordforms. This algorithm reads both versions of the same sentences, stores wordforms from each file and multi-wordform tokens form the converted version. It then aligns tokens using a robust synchronization algorithm that traverses both token sequences for a given sentence in a left-to-right manner. Whenever tokens do not match, the algorithm performs a lookahead on both token sequences until it is able to find a new anchor point, the "forward anchor". The search for a forward anchor is itself robust to tokenization mismatches, making use of the notion of "weak match" only used for comparing right contexts. The notion of weak match is defined as a disjunction of patterns; the main pattern looks for two consecutive matches or "pseudo-matches" between tokens in the original token sequence and tokens in the converted token sequence.[5] Once a forward anchor is found, tokens between the current position and the forward anchor are aligned according to a finite number of patterns, some of which are aware of the discrepancy in how some prepositions and determiners are agglutinated in the original tokenization scheme (e.g. *des* < *de les*).

Next, for each converted token which is aligned with an original token, its gold syntactic information is extracted from the original SPMRL annotations (`gold_SPMRL_head`, `gold_SPMRL_fpos`, `gold_SPMRL_delta`, `gold_SPMRL_label`) and associated with the converted token in the form of additional features, appended for convenience to the relevant field. These features respectively provide information about the head, fine-grained POS, distance from the governor and label of the current word's governor.

3.4. Parsing-based Treebank Correction

Inspired what had been tried when stacking a symbolic parser with DYALOG-SR (Villemonte de la Clergerie, 2014b), guiding gold features `pseudogold_UD_label` and `pseudogold_UD_delta` were added based on the result of the preliminary automatic conversion. They respectively refer in this preliminary UD version to the label and the (ordered) distance to the governor (if any). Obviously, with such features, which are not gold because of conversion errors but nevertheless quite accurate, training looks like a rather trivial task! However, based on a random process, about 6% of these guiding features were deleted, in order for the parser to learn how to correct a certain amount of errors, based on information about nearby dependencies, words, POS, and obviously SPMRL-based gold features added as per the previous section. It should be also noted that because all these feature are only indicative, the parser may even learn not to follow them under some con-

ditions, in other words, decide that some gold annotations are actually maybe not so correct.

Clearly, that kind of scheme (introducing a small amount of error) can not only be used to correct errors when converting to a new annotation schema (as tried here) but also to track and correct errors in gold annotations.

Initially developed for participating to the SPMRL shared task, the parser we used, DYALOG-SR, is a shift-reduce dependency parser, using Dynamic Programming and beams to explore its search space and a feature-rich perceptron to weight the parser actions (Villemonte De La Clergerie, 2013). Early and aggressive updates of the perceptron are performed at training time. In particular, following ideas from dynamic oracles (Goldberg and Nivre, 2012), updates may occur for actions that clearly results in violations of the gold tree, for instance when adding a bad dependency.

Using such a setting, our model was able to provide a high level of performance on the SEQUOIA gold data (10% not used in the training data and parsed with the same configuration as the data we aimed to correct) with 98.50% of LAS. The same range of accuracy was achieved on the dev and test section of the FTB that contained no conversion errors (resp. 98.48 and 98.64% of LAS).

4. Evaluation

Treebank conversion is a laborious task full of minutiae, and many conversion efforts improve their conversion in an iterative fashion, or as new relevant conversion needs are identified. A full manual evaluation of a converted treebank could represent an effort comparable to full re-annotation of a large part of the data. Indeed, few of the UD-conversion papers provide accuracy scores of the conversion on a manually annotated testbench.

For instance, The Danish conversion of Johannsen et al. (2015), uses a small set of hand-annotated sentences that reflect specific phenomena and hard cases that is used as held-out section during the iterative development of conversion rules. The Hungarian conversion of Vincze et al. (2017) uses a hand-corrected gold standard of 1,800 sentences. When comparing the quality of the conversion with the gold standard, they consider the accuracy (87.81 UAS and 75.99 LAS) not sufficient to release the resulting treebank.

We draw inspiration on their method to develop a hand-corrected sample to evaluate the quality of our conversion.One of the authors of the article, an expert in dependency annotation very familiar with the UD formalism, reviewed 100 sentences from the test section and 100 sentences from the dev section manually, correcting edges and labels that were either not properly attached, or not compliant with UD2.

Table 1 shows the scores for the manual validation. The Unlabeled Attachment Score (UAS) is very high, as the annotator did not disagree with most of the edges resulting from the conversion. However, the results are more drastic when analyzing the quality of the labels.

If we examine the label corrections by the expert annotation, we find that most of them reside on the label *fixed*, which has been used conservatively for all associated multiword expressions. Out of 360 relabelings overall, 274 are

---

[5]For instance a token *face* 'in front' in the converted token sequence will be considered as a pseudo-match with a token *face_à* 'in front of' in the original token sequence. This pseudo-match will result in an offset of 1 on the converted side, in order to skip the probable token *à* that follows the converted token *face*. A weak match will therefore be found if the converted token following this *à* is a match or pseudo-match with the token following *face_à* in the original token sequence.

Section	UAS	LAS
Dev	99.44	93.27
Test	99.40	94.75

Table 1: Manual evaluation scores for 100-sentences excerpts from the dev and test section.

relabelings for edges converted into *fixed* that should otherwise be *compound* or *flat:name*.

While some of the corrections for the *fixed* relation can be automated depending on the syntactic role of overall multiword subtree—e.g. a subtree that works as *case* is a multiword adposition and should be labeled *fixed*, while a *nsubj* label would per a proper name or a compound—the distinction between these tree types of relations, that are not exactly dependency relations in nature but must be described as such by virtue of the UD formalism, requires per-item linguistic analysis.

We have not observed any cases of mis-conversion of the core nominal arguments of verbs, which means that subjects and objects are always properly annotated, as well as the root note. In general, missattachments happen at lower points of the dependency tree that are closer to the leaves and are thus less relevant for overall dependency quality (Plank et al., 2015).

After multiword expressions, there are roughly thirty cases where the expert determined that the preferred relation should have been either *appos* (apposition) or *parataxis*. These are already controversial labels and are not easy to annotate. However, this indicates that the quality of the treebank is high enough for the most frequent expert relabelings to be within the domain of the fine distinctions of syntactic-semantic relations. Indeed, there was only one sentence out of the pooled 200 where there were present errors caused by coordination embedding, where the tree had to be corrected for the inner coordinates not to attach outside of the scope of their closest subsuming coordination.

5. Conclusion

We have described our effort to provide a highly reliable conversion of FTB into UD2.0 based on a convert-then-reparse principle. This method provides very high unlabeled accuracy (99.42 on average between 200 sentences). However, the quality of the resulting treebanks will need to be kept up to date with the advancements in the UD formalism, including a more homogeneous treatment of parataxis and appositions, as well as a detailed per-item analysis of multiword expressions and their potential relabeling. This method will be applied to the French Question Bank (Seddah and Candito, 2016) and to other data sets for English.

6. Acknowledgment

7. Bibliographical References

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

Bharat Ram Ambati, Tejaswini Deoskar, and Mark Steedman. 2013. Using ccg categories to improve hindi dependency parsing. In *ACL (2)*, pages 604–609.

Anders Björkelund, Ozlem Cetinoglu, Richárd Farkas, Thomas Mueller, and Wolfgang Seeker. 2013. (re) ranking meets morphosyntax: State-of-the-art results from the spmrl 2013 shared task. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 135–145.

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 169–172, Paris, France, October.

Marie Candito and Djamé Seddah. 2012. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proc. of TALN*.

Marie Candito, Benoit Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: Treebank conversion and first results. In *Proc. of LREC*.

Jan Einarsson. 1976. Talbankens talspråkskonkordans.

Richárd Farkas and Bernd Bohnet. 2012. Stacking of dependency and phrase structure parsers. In *COLING*, pages 849–866.

Yoav Goldberg and Joakim Nivre. 2012. A dynamic oracle for arc-eager dependency parsing. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 959–976.

Bruno Guillaume, Guillaume Bonfante, Paul Masson, Mathieu Morey, and Guy Perrier. 2012. Grew : un outil de réécriture de graphes pour le TAL. In *Proc. of TALN*, Grenoble, France.

Bruno Guillaume, Guy Perrier, and Guillaume Bonfante. (to appear). *Application de la Réécriture de Graphes au Traitement Automatique des Langues*.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal Dependencies for Danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza

Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phương Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Lilja Øvrelid, Jonas Kuhn, and Kathrin Spreyer. 2009. Improving data-driven dependency parsing using large-scale lfg grammars. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 37–40. Association for Computational Linguistics.

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *Eighteenth Conference on Computational Natural Language Learning (CoNLL 2015)*.

Corentin Ribeyre, Éric Villemonte De La Clergerie, and Djamé Seddah. 2015. Because Syntax does Matter: Improving Predicate-Argument Structures Parsing Using Syntactic Features. In *Proc. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, USA, June.

Djamé Seddah and Marie Candito. 2016. Hard time parsing questions: Building a questionbank for french. In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proc. of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA, October. Association for Computational Linguistics.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109.

Éric Villemonte De La Clergerie. 2013. Exploring beam-based shift-reduce dependency parsing with DyALog: Results from the SPMRL 2013 shared task. In *4th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL'2013)*.

Éric Villemonte De La Clergerie. 2014a. Jouer avec des analyseurs syntaxiques. In *TALN 2014*, Marseilles, France, July. ATALA.

Éric Villemonte de la Clergerie. 2014b. Jouer avec des analyseurs syntaxiques. In *Proc. of TALN 2014 (Volume 1: Long Papers)*, pages 67–78, Marseille, France, July. Association pour le Traitement Automatique des Langues.

Veronika Vincze, Katalin Simkó, Zsolt Szántó, and Richárd Farkas. 2017. Universal dependencies and morphology for hungarian-and on the price of universality. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 356–365.

Daniel Zeman, Filip Ginter, Jan Hajič, Joakim Nivre, Martin Popel, Milan Straka, and et al. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–20, Vancouver, Canada.