Cross-lingual Pronoun Prediction with Linguistically Informed Features

Rachel Bawden
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay,
91405 Orsay cedex, France
rachel.bawden@limsi.fr

Abstract

We present the LIMSI's cross-lingual pronoun prediction system for the WMT 2016 shared task. We use high-level linguistic features with explicit coreference resolution and expletive detection and rely on dependency annotations and a morphological lexicon. We show that our few, carefully chosen features perform significantly better than several language model baselines and competitively compared to the other systems submitted.

1 Introduction

This paper describes the LIMSI's submission to the cross-lingual pronoun prediction shared task at WMT 2016 (Guillou et al., 2016) for the language direction English to French. The task involves classifying the subject pronouns *it* and *they* into the French pronoun classes *il*, *ils*, *elle*, *elles*, *ce*, *cela*, *on* and *OTHER* (which also includes the null pronoun). Target sentences are human translations, in which pronouns to be predicted are replaced by placeholders. An automatic word alignment is given between English and French sentences. Unlike the same version of the task for DiscoMT 2015 (Hardmeier et al., 2015), target sentences are supplied in lemmatised and part-of-speech (PoS) tagged format, without the original tokens.[1] The official metric for the task is the macro-averaged recall, which has the effect of giving more weight to rarer pronouns. Training data is news and speech-based and the development and test sets are speech transcriptions (Ted Talks).

Our system is based on a statistical feature-based classification approach. It is linguistically motivated with carefully chosen, high-level features designed to tackle particular difficulties of the classification problem, including explicit anaphora resolution using coreference chains and the detection of expletive pronouns.

On top of a set of language model-based features, which form our baseline, we design a set of features to exploit linguistic annotations and resources for: (i) coreference resolution and expletive detection to guide the prediction of the pronoun classes *il*, *ils*, *elle* and *elles*, (ii) local context features based on syntactic dependencies, and (iii) the use of highly discriminative corpus-extracted contexts, in particular for the *OTHER* class.

2 Linguistic challenges of the task

There are a number of difficulties in the translation of the subject pronouns *it* and *they* into French. A major issue is that, in French, pronouns and nouns are marked for grammatical gender (masculine and feminine) and number (singular and plural), whilst in English, *it* and *they* are only marked for number. When French pronouns are anaphoric, (i.e. they refer to an entity that is present in the text or context), their gender and number is almost always determined by their referent.[2] Knowing which pronoun to use therefore relies on knowing to which noun the pronoun refers as well as the gender and number of the noun. Automatic tools exist for anaphora resolution, often also constructing coreference chains to link all mentions that refer to the same entity. PoS tags and morphological lexica can provide information about gender and number. This is of course a simplification,

[1] In many cases the morphology of the surrounding local context could supply the correct pronoun. Not a single submission scored higher than the language model baseline according to the official metric, the macro-averaged F-score.

[2] There are some exceptions, such as the singular, gender-neutral *they*. Another example is when the referential expression refers to a group of people, such as *équipe* 'team'. The anaphoric pronoun can be a plural *ils* 'they' rather than singular. Common in English, and although less accepted in French, there exist several examples of this in the task data.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 564–570,
Berlin, Germany, August 11-12, 2016.

and the situation is in reality much more complex, for example when the referent is two coordinated nouns or when the English pronoun is the *singular*, gender-neutral pronoun *they*. There is also the case of the indefinite pronoun *on*, which is used as a translation of the indefinite English pronoun *one, you*, and sometimes *they*.

An added difficulty is the fact that *it* is sometimes translated as the expletive (or impersonal) *il*, as in ***il*** *pleut* '**it** is raining'. These should not be confused with the anaphoric pronouns, and not all automatic coreference tools explicitly detect them. Dependency parsing can be particularly useful for detecting them via individual local features, such as looking at the verb on which the pronoun depends. There are also other possible translations of *it*, namely *ce* and the demonstrative pronoun *cela/ça*, which can sometimes be predicted from the context, but are often difficult to translate.

In the task data, the English pronoun is frequently aligned with a word that does not belong to the 7 main pronoun classes described above, or is simply not translated at all. In these cases, the target pronoun is said to belong to the class *OTHER*, a class that is frequent, heterogeneous and therefore likely to pose problems for prediction.

3 System overview

To resolve these difficulties, we choose to privilege the use of linguistic tools and resources to exploit a small number of linguistically motivated features rather than approach the problem by using a great number of weakly motivated features.

3.1 Tools and resources

We used various annotations for both English source sentences and French target sentences: PoS tagging and dependency parsing for both languages, coreference resolution for English and morphological analysis for French. English annotations were all produced using the Stanford CoreNLP toolkit (Manning et al., 2014). Standard, pretrained parsing models could not be used on the lemma-based French sentences, and we therefore re-trained a parsing model solely based on lemmas and PoS-tags, using the Mate Graph-based transition parser (Bohnet and Nivre, 2012) and the French training data for the 2014 SPMRL shared task (Seddah et al., 2014). Some pre-processing was necessary to create a compatible tagset between the SPMRL data and the task training data.[3] We enriched the French annotations using a morphological and syntactic lexicon, the Le*fff* (Sagot, 2010), to include noun gender by mapping lemmas to their genders (allowing for ambiguity). We also used the lexicon to provide information about impersonal verbs and adjectives (Sec. 3.2.2).

3.2 Linguistic features

We use as our main baseline a set of language model features (Sec. 3.2.1), which also form the starting point of our system. We add to this three types of features: coreference resolution and expletive detection (Sec. 3.2.2), local, syntax-based features (Sec. 3.2.3) and a syntactic context template feature (Sec. 3.2.4).

3.2.1 Language model features

Using a language model provides a way of modelling local context using the words immediately surrounding the pronoun. In our case, it provides no information concerning number, since the French target sentences are lemmatised, and the feminine gender is also unlikely to be well predicted by the model in the case of anaphoric pronouns unless the referent is in a very local context.

We base our language model features on the pronoun class probabilities provided by the task organisers as part of the official language model baseline. These features are based on the probability of the most probable pronoun class as per the language model: (i) the most probable class, (ii-iv) the most probable class if its probability is superior to 90%, 80%, 50%, and (v) the concatenation of the two most probable classes.

3.2.2 Coreference features

We use two features to represent anaphora resolution, namely the gender (masculine, feminine or impersonal) and number (singular, plural or impersonal) of the pronoun's referent.

Standard anaphora resolution: To identify the referent of an anaphoric pronoun, we applied the Stanford coreference resolver (de Marneffe et al., 2015) to the English sentences, separated by document, and used the automatic alignments to identify the corresponding referent in French (see Figure 1). Gender is determined by that of the French referent (as provided by the Le*fff*). Since French

[3] We analysed the quality of the syntactic annotations, using the SPMRL test set and scorer, to give an unlabelled attachment score of 89.83%.

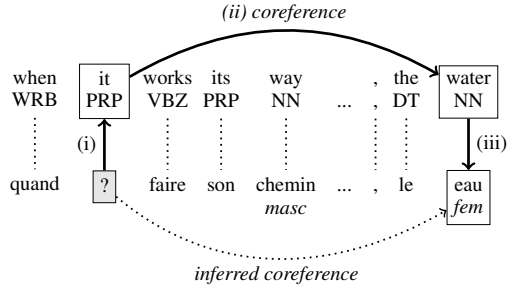

Figure 1: Use of coreference chains to determine gender and number of anaphoric pronouns.

sentences are lemmatised, number must be sought in the English sentence. We test two variants, in which number is determined by (i) the number of the English referent (which is integrated in the PoS tagset), as shown in Figure 1, and (ii) the number of the aligned English pronoun: singular for *it* and plural for *they*. Coreference chains can cross sentence boundaries, and mentions can span several words, in which case we took information associated with the mention's head.

The accuracy of our coreference features depends on the ability of the coreference tool to detect accurate and complete chains, the quality of the automatic alignments, the accuracy of the PoS tags to predict number and the coverage of the lexicon for French noun gender.

We evaluate the quality of the coreference tool on the development set by manually annotating the French pronouns and comparing the predicted and gold referents. Of 237 pronouns of the form *il*, *elle*, *ils* or *elles*, 194 were anaphoric with a textual referent. The correct coreferent was provided in only 52.6% of cases, the majority being for the masculine plural class *ils*. Moreover, 32% of these pronouns were linked only to other pronouns, therefore with no explicit referent (in particular for the feminine plural *elles*). The tool also often fails to predict impersonal pronouns, erroneously supplying coreference chains for 18 impersonal pronouns out of 25.

Back-off anaphora resolution: Given these insufficiencies of the coreference tool, we developed a back-off coreference method, in cases where it provides no gender and number. It consists of providing additional values for the two coreferences features by taking the nearest preceding noun phrase in the previous sentence as the pronoun's referent. Although likely to add a certain amount of noise, especially in cases where the pronoun is non-anaphoric, this method provides more data values.

Expletive pronoun detection: One case of non-anaphoric pronoun detection that can be dealt with directly is the case of the French impersonal pronoun *il*. We apply heuristic rules[4] to detect such impersonals on the French side, modifying the coreference feature values to *impersonal* when one is detected. We consider a pronoun to be an impersonal *il* when it is in an impersonal construction (containing an impersonal verb or adjective), information provided by a look-up in the Le*fff*. Certain cases of non-ambiguous impersonals such as *il* ***faut*** *le faire* 'it must be done' are easily dealt with. Ambiguous cases, where the adjective or verb can be used both personally and impersonally, can be disambiguated by the context, for example by the presence of a following *de* 'to' for verbs and adjectives or *que* 'that' for verbs.[5]

3.2.3 Local features

For the other pronouns, *ce*, *cela*, *on* and *OTHER*, the local context plays a crucial role. We include a number of local, syntax-guided context features, based on the syntactic governor, as provided by the dependency parse. The features include the form of the English aligned token (raw and lowercased), the form, PoS tag and lemma of the syntactic governor of the English aligned token and the PoS tag and lemma of the syntactic governor of the French pronoun. Finally, we include a boolean feature indicating whether or not the pronoun is found at the beginning of the sentence.

3.2.4 Context template feature

We also look at the target pronoun's wider and richer context, using relative and syntactic positions, to produce a single, strong feature, whose value is the class (if any) to which the pronoun's context indicates that it is particularly likely to be associated. In a preliminary step, we extracted all context templates from the training and development sets defined by storing the lemmas and PoS tags of the words at the following positions: (i) 2 following, (ii) 1 preceding and 2 following, (iii) 1

[4]Tools do exist for impersonal detection, however they are designed to process tokens and not lemmas.

[5]For example, *il est intéressant.* 'it/he is interesting' vs. *il est intéressant de…* 'it is interesting to…'

Position Relative to the pronoun							
-1	+1	+2	+3	gov.	class	Num.	%
	un	NOM			OTHER	1503	99
VER				NOM_{det}	OTHER	1003	97
la/le				VER_{subj}	on	478	96
,	être	ADJ	que		il	4131	98
PUN	être	ADJ	de		il	5239	95

Table 1: Examples of context templates with their associated class. We also give the percentage of occurrences of the template with the associated class and their frequency of co-occurrence.

preceding and 3 following, (iv) the governor, (v) the governor and the function, (vi) the governor and its governor, and (vii) the preceding token and the governor and its function.

See Table 1 for some examples of context template values, linked with a certain class, for which they are particularly well associated. This is indicated by the high frequency of occurrence of the <template, class> pair and the high percentage of occurrences of the template with the class, as observed in the training and development sets.

Relevance score used: Our aim was to select the pairs that were the most discriminative for the corresponding class and which were most frequent, in order to create an aggregated, reliable feature. We therefore ranked the pairs according to the following heuristic relevance score based on frequency counts in the corpora (Equation 1).

$$score(<c,y>) = \frac{\text{occ}(<c,y>)}{\sum_{y' \in Y} \text{occ}(<c,y'>)} \sqrt{\text{occ}(<c,y>)} \quad (1)$$

where *c* is a given context, *y* a given class and *Y* is the set of possible classes.

The score is designed to be a reasonable compromise between the probability of the context being associated with the given class and their frequency of co-occurrence.[6] We select the 10,000 top-ranked pairs and further filter to only keep pairs where the context is associated with the class more than 95% of the time.[7] When the pronoun to be predicted is found within the context of one of these templates, the feature value is the class associated with the context. A total of 5,003 templates were retained: 2,658 for *OTHER*, 1,987 for *il*, 347 for *ce*, 9 for *on* and 2 for *cela*.

The templates are particularly useful for detecting the *OTHER* class, which include empty instances (where the English pronoun is untranslated) and words other than the 7 target pronoun classes. For example, if followed by the determiner *un* and a noun, there is a strong association with the *OTHER* class (first example in Table 1). They can be especially useful in cases of alignment problems or anomalous predictions, and also for detecting certain collocations.

3.3 Classification setup

We use a random forest classifier, as implemented in `Scikit-learn` (Pedregosa et al., 2011). Our choice of machine learning algorithm is partly based on the ability of random forests to account for class imbalance and outliers, a necessary trait in the case of this task.[8] They also have the advantage of not being linear, and therefore of being able to find patterns in the data using a relatively small number of features, as is our aim here.[9] We split the task into separate classifiers for *it* and *they*; a preliminary comparative study suggested that this produces slightly better results than training a single classifier for all source pronouns.

4 Results

We provide the results of several variants of our system, in order to analyse the different components. We report scores for the two official baselines $baseline_{WMT\text{-}1}$ and $baseline_{WMT\text{-}2}$. We also provide two extra baselines: $baseline_{mostFreqPro}$, which predicts the most frequent class for each English pronoun (masc. sg. *il* for *it* and masc. pl. *ils* for *they*) and a second, $baseline_{LM}$, which uses as features the form of the English pronoun (*it* or *they*) and the language model features described in Sec. 3.2.1. All scores are produced using the official evaluation script and are reported "as is" using two significant decimal figures.

A minor implementation issue was found concerning the use of the context templates for the two submissions. We nevertheless include the results

[6] Although not normalised, the score, which is greater for a more relevant pair, has the advantage of being constant for a given probability and frequency count, and is therefore not dependent on the rarity of either the class or the context, unlike similar measures such as the log-likelihood ratio.

[7] We tested several values in preliminary experiments on the development set and found these values to be a good compromise between score optimisation and training time.

[8] Please refer to the Shared Task overview paper for the class distributions.

[9] We use Gini as the optimising criterion, 250 estimators, a maximum depth of 500 and a minimum number of leaf samples of 1. All other parameters are those provided by default.

System	Macro-avg. Recall (%) Dev	Test	Acc. (%) Test
$\text{baseline}_{\text{WMT-1}}$	40.63	46.98	52.01
$\text{baseline}_{\text{WMT-2}}$	-	50.85	53.35
$\text{baseline}_{\text{mostFreqPro}}$	24.03	24.39	34.58
$\text{baseline}_{\text{LM}}$	48.63	55.21	65.95
*LIMSI_1	56.14	59.32	68.36
*LIMSI_2	55.08	59.34	68.36
LIMSI_1	55.65	60.94	69.44
LIMSI_2	54.82	59.37	68.36
$\text{LIMSI}_{1,\text{NoLM}}$	51.66	54.35	62.73
$\text{LIMSI}_{2,\text{NoLM}}$	50.87	54.94	63.54
$\text{LIMSI}_{1,\text{SimpleCR}}$	55.45	61.26	71.05
$\text{LIMSI}_{2,\text{SimpleCR}}$	**56.16**	60.58	70.51

Table 2: Comparative results of baseline systems, the LIMSI submissions and several variants.

of these two systems (marked with an asterisk), whose results do not however differ wildly from those of the corrected versions. The two different versions (labelled 1 and 2) correspond to the two different methods of providing the number value of the coreference features (see Sec. 3.2.2): the first method taking the number of the last referent identified by the coreference tool, and the second from the form of the aligned English pronoun.

We provide two additional variants for each version. *NoLM* variants do not use language model features, whereas *SimpleCR* variants only rely on the Stanford tool for coreference resolution, excluding our back-off method (see Sec. 3.2.2).

5 Discussion

The evaluation metric for the task (macro-averaged recall) is such that very sparse classes hold a huge weight in the final evaluation.[10] There are also vast differences in classification quality between the datasets, as illustrated by the systematic percentage point increase in score (up to 6 points) between the development and the test set. This highlights the fact that the heterogeneity of data should be taken into account when designing a system, and supports the idea of features based on external (and therefore static) linguistic resources rather than relying too much on the data itself. The result is that our best performing system during development is not always our best performing on the test set (see the results of $\text{LIMSI}_{1,\text{SimpleCR}}$ vs. $\text{LIMSI}_{2,\text{SimpleCR}}$).

[10]Correctly predicting a single extra *on* improves the overall score by more than 1%.

There is no significant difference between the two variants of the LIMSI system. However the first variant performs better on both development and test sets more often than the second.

Compared to the four baselines, the linguistically rich systems perform systematically better. The much lower scores of $baseline_{LM}$ compared to $LIMSI_1$ and $LIMSI_2$ show that adding our linguistic features provides extra and different information from the language model features. A slightly disconcerting observation is that if we remove the language model features ($LIMSI_{1,NoLM}$ and $LIMSI_{2,NoLM}$), the score compared to $baseline_{LM}$ is up to 3 percentage points higher on the development set, but lower on the test set, suggesting that the information needed to predict the pronouns in the test set was probably mostly local, requiring less linguistic knowledge, another effect of the different natures of the sets and their small sizes.

The experiments with simple coreference give comparable scores on the development set and higher scores on the test set (up to 61.26% macro-averaged recall for $LIMSI_{1,SimpleCR}$). It is difficult to draw any conclusions about which method of gender and number induction is best, although our back-off method appears to be too noisy.

5.1 Finer analysis

The classification matrix for the results on the test set for $\text{LIMSI}_{2,\text{SimpleCR}}$ (the best performing model on the development set) is shown in Table 3. Unsurprisingly, the most problematic classes are *elle* and *elles*, for which the only means of correctly predicting the gender is to have access to the pronoun's textual referent and its gender. Although a majority of the feminine pronouns were classified as having the correct number, only 3 out of 25 occurrences of *elles* were assigned the correct class. The other two classes for which the system performed less well were *cela* (often confused with *il*) and *on* (confused with *ils* and *OTHER*). These were all the least frequent pronoun classes, which therefore have a large impact on the overall score because of the macro-averaged metric. The classes which were best predicted were *ce*, with a high precision of 91.53%, *OTHER* with a high recall of 88.24% and *ils* with a recall of 78.87%.

5.2 Oracle coreference resolver

One of the weaknesses of the system is, as expected, the prediction of the gender of the French pronoun, which is dependent on the quality of an

	Classified as											
	ce	elle	elles	il	ils	cela	on	other	SUM	P (%)	R (%)	F (%)
ce	**54**	1	0	11	0	0	0	2	68	91.53	79.41	85.04
elle	0	**13**	1	6	0	2	0	1	23	41.94	56.52	48.15
elles	1	2	**3**	1	13	1	0	4	25	23.08	12.00	15.79
il	2	7	0	**44**	1	2	1	4	61	61.97	72.13	66.67
ils	0	1	9	0	**56**	0	0	5	71	75.68	78.87	77.24
cela	0	5	0	7	0	**13**	1	5	31	72.22	41.94	53.06
on	0	0	0	0	2	0	**5**	2	9	55.56	55.56	55.56
OTHER	2	2	0	2	2	0	2	**75**	85	76.53	88.24	81.97
SUM	59	31	13	71	74	18	9	98				
Micro-averaged										70.51	70.51	70.51
Macro-averaged										62.31	**60.58**	60.43

Table 3: A decomposition of results for the system $\text{LIMSI}_{2,\text{SimpleCR}}$ on the test set.

external coreference tool. In order to assess the performance of our system independently of this specific tool, we imagine a scenario in which we have access to perfect impersonal detection and coreference resolution and can therefore correctly predict all instances of *il*, *ils*, *elle* and *elles*. This gives perfect recall for these four pronouns and enables us to assess the capacity of the system's other features to distinguish between the remaining pronouns, had coreference resolution been perfect.

We first automatically detect the impersonal pronoun *il* using the dedicated tool *ilimp* (Danlos, 2005). Since the tokenised French sentences were available for the French-to-English version of the same task, we directly applied the tool to raw training and development sentences. For the remaining personal pronouns, we take gender and number directly from the gold label, as if a coreference system had correctly predicted them.

The results (for the development set) when using oracle coreference resolution, with a macro-averaged recall of 85.31%, show that if the anaphoric pronouns are predicted with 100% precision and recall, there are still lacunas in the system, notably for the label *on*, for which the precision is 57.14% and the recall only 40%, due to 6 out of 10 occurrences being classified as *OTHER*. The other class with a low recall (although a high precision of 97.14%) is *cela*, for which 25 out of 63 occurrences were incorrectly classified as *OTHER*. This suggest that there is a positive bias towards the *OTHER* class, which is the third most frequent. We speculate that the overprediction of this class could be due to the context template feature, which was geared to predict the *OTHER* class. Having such a statistically strong feature, with contexts highly related to a certain class does not allow for exceptions to the rule.

This shows that there is room for improvement for the other pronouns, even with perfect coreference resolution. To improve the use of context templates, there are two options. Firstly, the thresholds for the inclusion of templates could be revised; they could either be increased to reinforce the feature's strength, or decreased to allow for more noise, enabling other features to counterbalance it in some cases. Secondly, more well-designed features that allow for a greater decomposition of decisions could be used, rather than relying on a single feature that does not allow any deviation from the rule.

6 Conclusion

We have presented a linguistic, feature-based pronoun prediction system, using explicit anaphora resolution and expletive detection. We have explored the use of dependencies for local context features and discriminative context templates to target particular difficulties of the task. Our results are well above the baseline, and our system was ranked sixth out of nine submissions. We see two possible improvements for the system, either relying on a more sophisticated, better performing language model (such as LSTMs), or, more interestingly, improving our linguistic features and the resources and tools that they are based on.

The approach is generalisable to other language pairs, provided that similar tools and resources are available for those languages. The features would have to be adjusted to take into account the different pronoun mappings of the two languages. For example, for the reverse direction, French to English, named entities and animacy features are crucial for mapping the French pronouns *il/elle* to *s/he* for gender-specific beings such as people and to *it* for objects.

References

Bernd Bohnet and Joakim Nivre. 2012. A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*, pages 1455–1465, Jeju Island, Korea.

Laurence Danlos. 2005. Automatic recognition of French expletive pronoun occurrences. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP '05)*, pages 73–78, Jeju Island, Korea.

Marie-Catherine de Marneffe, Marta Recasens, and Christopher Potts. 2015. Modeling the Lifespan of Discourse Entities with Application to Coreference Resolution. *Journal of Artificial Intelligence Research*, 52:445–475.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the 1st Conference on Machine Translation (WMT '16)*, Berlin, Germany.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation (DiscoMT '15)*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '14)*, pages 55–60, Baltimore, Maryland, USA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Benoît Sagot. 2010. The Le*fff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, pages 2744–2751, Valletta, Malta.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the 1st Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages (SPMRL-SANCL '14)*, pages 103–109, Dublin, Ireland.