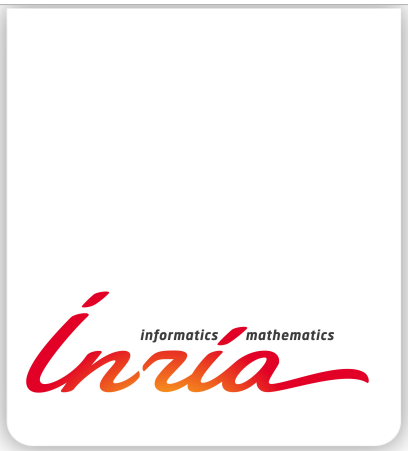

# External Lexical Information for Multilingual Part-of-Speech Tagging

**Benoît Sagot**

**RESEARCH REPORT**
**N° 8924**
June 2016
Project-Teams ALPAGE

ISSN 0249-6399 ISRN INRIA/RR--8924--FR+ENG




# External Lexical Information for Multilingual Part-of-Speech Tagging

Benoît Sagot

Project-Teams ALPAGE

Research Report n° 8924 — June 2016 — 15 pages

**Abstract:** Morphosyntactic lexicons and word vector representations have both proven useful for improving the accuracy of statistical part-of-speech taggers. Here we compare the performances of four systems on datasets covering 16 languages, two of these systems being feature-based (MEMMs and CRFs) and two of them being neural-based (bi-LSTMs). We show that, on average, all four approaches perform similarly and reach state-of-the-art results. Yet better performances are obtained with our feature-based models on lexically richer datasets (e.g. for morphologically rich languages), whereas neural-based results are higher on datasets with less lexical variability (e.g. for English). These conclusions hold in particular for the MEMM models relying on our system MElt, which benefited from newly designed features. This shows that, under certain conditions, feature-based approaches enriched with morphosyntactic lexicons are competitive with respect to neural methods.

**Key-words:** Part-of-Speech Tagging, Feature-based models, Neural models, MEMM, CRF, bi-LSTM, Multilingual Analysis

RESEARCH CENTRE
PARIS – ROCQUENCOURT

Domaine de Voluceau, - Rocquencourt
B.P. 105 - 78153 Le Chesnay Cedex

# Utilisation d'informations lexicales externes pour l'annotation multilingue en parties du discours

**Résumé :** Les lexiques morphosyntaxiques et les représentations vectorielles des mots ont chacun montré leur utilité pour améliorer la précision d'étiqueteurs morphosyntaxiques statistiques. Nous comparons ici les performances de quatre systèmes sur des jeux de données couvrant 16 langues, deux de ces systèmes reposant sur des traits (MEMM et CRF) et deux autres sur des approches neuronales (bi-LSTM). Nous montrons qu'en moyenne les quatre approches obtiennent des performances similaires de niveau état-de-l'art. Néanmoins, nos modèles reposant sur des traits ont de meilleures performances sur les jeux de données lexicalement plus riches (par exemple sur des langues à morphologie riche), alors que les résultats obtenus par les approches neuronales sont meilleurs sur les jeux de données dont la variabilité lexicale est moindre (par exemple pour l'anglais). Ces conclusions sont vraies en particulier pour nos modèles de type MEMM faisant usage de notre système MElt, qui s'appuie sur un jeu de traits renouvelé. Ceci montre que, sous certaines conditions, les approches par traits enrichies par des lexiques morphosyntaxiques sont compétitifs par rapport aux approches neuronales.

**Mots-clés :** Étiquetage en partie du discours, Modèles reposant sur des traits, Modèles neuronaux, MEMM, CRF, bi-LSTM, Analyse multilingue

# 1 Introduction

Part-of-speech tagging is now a classic task in natural language processing, for which many systems have been developed or adapted for a large variety of languages. Its aim is to associate each "word" with a morphosyntactic tag, whose granularity can range from a simple morphosyntactic category, or part-of-speech (hereafter PoS), to finer categories enriched with morphological features (gender, number, case, tense, mood, etc.).

The use of machine learning algorithms trained on manually annotated corpora has long become the standard way to develop PoS taggers. A large variety of algorithms have been used, such as (in approximative chronological order) bigram and trigram hidden Markov models (Merialdo, 1994; Brants, 1996, 2000), decision trees (Schmid, 1994; Magerman, 1995), maximum entropy Markov models (MEMMs) (Ratnaparkhi, 1996) and Conditional Random Fields (CRFs) (Lafferty et al., 2001; Constant and Tellier, 2012). With such machine learning algorithms, it is possible to build PoS taggers for any language, provided adequate training data is available.

As a complement to annotated corpora, it has previously been shown that external lexicons are valuable sources of information, in particular morphosyntactic lexicons, which provide a large inventory of (word, PoS) pairs. Such lexical information can be used in the form of constraints at tagging time (Kim et al., 1999; Hajič, 2000) or during the training process as additional features combined with standard features extracted from the training corpus (Chrupała et al., 2008; Goldberg et al., 2009; Denis and Sagot, 2012).

In recent years, a different approach to modelling lexical information and integrating it into natural language processing systems has emerged, namely the use of vector representations for words or word sequences (Bengio et al., 2003; Collobert and Weston, 2008; Chrupała, 2013; Ling et al., 2015; Ballesteros et al., 2015; Müller and Schütze, 2015). Such representations, which are generally extracted from large amounts of raw text, have proved very useful for numerous tasks including PoS tagging, in particular when used in recurrent neural networks (RNNs) and more specifically in mono- or bi-directional, word-level and/or character-level long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997; Ling et al., 2015; Ballesteros et al., 2015; Plank et al., 2016).

Both approaches to representing lexical properties and to integrating them into a PoS tagger improve tagging results. Yet they rely on resources of different natures. The main advantage of word vectors is that they are built in an unsupervised way, only requiring large amounts of raw textual data. They also encode finer-grained information than usual morphosyntactic lexicons, most of which do not include any quantitative data, not even simple frequency information. Conversely, lexical resources often provide information about scarcely attested words, for which corpus-based approaches such as word vector representations are of limited relevance. Moreover, morphological or morphosyntactic lexicons already exist for a number of languages, including less-resourced langauges for which it might be difficult to obtain the large amounts of raw data necessary to extract word vector representations.

Our main goal is therefore to compare the respective impact of external lexicons and word vector representations on the accuracy of PoS models. This question has already been investigated for 6 languages by Müller and Schütze (2015) using the state-of-the-art CRF-based tagging system MarMoT. The authors found that their best-performing word-vector-based PoS tagging models outperform their models that rely on morphosyntactic resources (lexicons or morphological analysers). In this paper, we report on larger comparison, carried out in a larger multilingual setting and comparing different tagging models. Using different 16 datasets, we compare the performances of two feature-based models enriched with external lexicons and of two LSTM-based models enriched with word vector representations. A secondary goal of our work is to compare the relative improvements linked to the use of external lexical information in the two

feature-based models, which use different models (MEMM vs. CRF) and feature sets.

More specifically, our starting point is the MElt system (Denis and Sagot, 2012), an MEMM tagging system. We first briefly describe this system and the way we adapted it by integrating our own set of corpus-based and lexical features. We then introduce the tagging models we have trained for 16 different languages using our adapted version of MElt. These models are trained on the *Universal Dependencies* (v1.2) corpus set (Nivre and al., 2015), complemented by morphosyntactic lexicons. We compare the accuracy of our models with the scores obtained by the CRF-based system MarMoT (Müller et al., 2013; Müller and Schütze, 2015), retrained on the same corpora and the same external morphosyntactic lexicons. We also compare our results to those obtained by the best bidirectional LSTM models described by Plank et al. (2016), which both make use of Polyglot word vector representations published by Al-Rfou et al. (2013). We will show that an optimised enrichment of feature-based models with morphosyntactic lexicon results in significant accuracy gains. The macro-averaged accuracy of our enriched MElt models is above that of enriched MarMoT models and virtually identical to that of LSTMs enriched with word vector representations. More precisely, per-language results indicate that lexicons provide more useful information for languages with a high lexical variability (such as morphologically rich languages), whereas word vectors are more informative for languages with a lower lexical variability (such as English).

## 2 MElt

MElt (Denis and Sagot, 2012) is a tagging system based on maximum entropy Markov models (MEMM) (Ratnaparkhi, 1996), a class of discriminative models that are suitable for sequence labelling (Ratnaparkhi, 1996). The basic set of features used by MElt is given in (Denis and Sagot, 2012). It is a superset of the feature sets used by Ratnaparkhi (1996) and Toutanova and Manning (2000) and includes both *local standard features* (for example the current word itself and its prefixes and suffixes of length 1 to 4) and *contextual standard features* (for example the tag just assigned to the preceding word). In particular, with respect to Ratnaparkhi's feature set, MElt's basic feature set lifts the restriction that local standard features used to analyse the internal composition of the current word should only apply to rare words.

One of the advantages of feature-based models such as MEMMs and CRFs is that complementary information can be easily added in the form of additional features. This was investigated for instance by Kübler et al. (2010), whose best-performing model for PoS tagging dialogues was obtained with a version of MElt extended with dialogue-specific features. Yet the motivation of MElt's developers was first and foremost to investigate the best way to integrate lexical information extracted from large-scale morphosyntactic lexical resources into their models, on top of the training data (Denis and Sagot, 2012). They showed that performances are better when this external lexical information is integrated in the form of additional *lexical features* than when the external lexicon is used as constraints at tagging time.[1] These lexical features can also be divided into *local lexical features* (for example the list of possible tags known to the external lexicon for the current word) and *contextual lexical features* (for example the list of possible tags known to the external lexicon for surrounding words). In particular, lexical contextual features provide a means to model the right context of the current word, made of words that have not yet been tagged by the system but for which the lexicon often provides a list of possible tags. Moreover, tagging accuracy for out-of-vocabulary (OOV) words is improved, as a result of the fact that words unknown to the training corpus might be known to the external lexicon.

[1] For instance by constraining the tagger in such a way that words known to the lexicon can only be associated with tags provided by the lexicon.

| | | |
|---|---|---|
| | **Local standard features** | |
| | $\text{wd} = w_i$ | $\wedge\ t_i = T$ |
| $\forall k \in [1..4]$ | $\text{pref}^k = w_i^1 \dots w_i^k$ | $\wedge\ t_i = T$ |
| $\forall k \in [1..5]$ | $\text{suff}^k = w_i^{n_i-k+1} \dots w_i^{n_i}$ | $\wedge\ t_i = T$ |
| | $\text{nb} = \text{containsDigit}(w_i)$ | $\wedge\ t_i = T$ |
| | $\text{hyph} = \text{containsHyphen}(w_i)$ | $\wedge\ t_i = T$ |
| | $\text{uc} = \text{containsUppercase}(w_i)$ | $\wedge\ t_i = T$ |
| | $\text{niuc} = (\text{containsUppercase}(w_i) \wedge i > 1)$ | $\wedge\ t_i = T$ |
| | $\text{auc} = \text{containsOnlyUppercase}(w_i)$ | $\wedge\ t_i = T$ |
| | **Contextual standard features** | |
| | $\text{wd}_{-2} = w_{i-2}$ | $\wedge\ t_i = T$ |
| | $\text{wd}_{-1} = w_{i-1}$ | $\wedge\ t_i = T$ |
| | $\text{wd}_{+1} = w_{i+1}$ | $\wedge\ t_i = T$ |
| | $\text{wd}_{+2} = w_{i+2}$ | $\wedge\ t_i = T$ |
| | $\text{swds} = w_{i-1}.w_{i+1}$ | $\wedge\ t_i = T$ |
| $\forall k \in [1..3]$ | $\text{pref}_{+1}^k = w_{i+1}^1 \dots w_i^k$ | $\wedge\ t_i = T$ |
| $\forall k \in [1..3]$ | $\text{suff}_{+1}^k = w_{i+1}^{n_i-k+1} \dots w_i^{n_i}$ | $\wedge\ t_i = T$ |
| | $\text{ptag}_{-2} = t_{i-2}$ | $\wedge\ t_i = T$ |
| | $\text{ptag}_{-1} = t_{i-1}$ | $\wedge\ t_i = T$ |
| | $\text{ptags} = t_{i-2}.t_{i-1}$ | $\wedge\ t_i = T$ |
| | **Local lexical features** | |
| if $\text{lex}(w_i) = \{t\}$ | $\text{lex}_{\text{u}} = t$ | $\wedge\ t_i = T$ |
| if $\lvert\text{lex}(w_i)\rvert > 1, \forall t_j \in \text{lex}(w_i)$ | $\text{lex}_{\text{in}} = t_j$ | $\wedge\ t_i = T$ |
| if $\lvert\text{lex}(w_i)\rvert > 1$ | $\text{lex}_{\text{disj}} = \bigvee_{t_j \in \text{lex}(w_i)} t_j$ | $\wedge\ t_i = T$ |
| | **Contextual lexical features** | |
| | $\text{lex}_{+1} = \bigvee_{t_j \in \text{lex}(w_{i+1})} t_j$ | $\wedge\ t_i = T$ |
| | $\text{lex}_{+2} = \bigvee_{t_j \in \text{lex}(w_{i+2})} t_j$ | $\wedge\ t_i = T$ |
| | $\text{lex}_{+1.2} = \left(\bigvee_{t_j \in \text{lex}(w_{i+1})} t_j\right) . \left(\bigvee_{t_j \in \text{lex}w_{i+2}} t_j\right)$ | $\wedge\ t_i = T$ |
| | **Contextual hybrid features** | |
| | $\text{ptag}_{-1}.\text{lex}_{+1} = t_{i-1}.(\bigvee_{t_j \in \text{lex}(w_{i+1})} t_j)$ | $\wedge\ t_i = T$ |

Table 1: Feature set used by our MElt models. The current word is $w_i = w_i^1 \dots w_i^{n_i}$. Previously assigned tags for the two previous words are $t_{i-2}$ and $t_{i-1}$. The tag to be predicted for the current word is $t_i$, which can be assigned any tag $T$ in the tagset. The lex function applied to a word returns the set of all tags known to the lexicon for this word, or the singleton {_unk_} if the word is unknown to the lexicon. Boolean functions used by the local standard features have self-explanatory names.

| Language | Source Lexicon | #entries | tagset size | Reference |
|---|---|---|---|---|
| Bulgarian | Multext-EAST | 53056 | 12 | (Erjavec, 2010) |
| Croatian | HML | 1360687 | 22 | (Oliver and Tadić, 2004) |
| Czech | Morfflex (parts) | 2094860 | 65 | (Hajič and Hlaváčová, 2013) |
| Danish | STO | 566184 | 13 | (Braasch et al., 2008) |
| English | EnLex | 472850 | 22 | (Sagot, 2010) |
| French | Le*fff* | 539278 | 25 | (Sagot, 2010) |
| German | DeLex | 465434 | 52 | (Sagot, 2014) |
| Indonesian | Kateglo | 72217 | 118 | `https://github.com/ivanlanin/kateglo` |
| Italian | Morph-it! | 422756 | 31 | (Zanchetta and Baroni, 2005) |
| Norwegian | OrdBank | 679763 | 19 | (Hagen and Nøklestad, 2010) |
| Persian | PerLex | 511840 | 37 | (Sagot and Walther, 2010) |
| Polish | PolLex | 390370 | 28 | (Sagot, 2007) |
| Portuguese | LABEL-LEX$_{SW}$ | 971514 | 29 | (Ranchhod et al., 1999) |
| Slovenian | SloLeks | 957525 | 25 | (Krek et al., 2008) |
| Spanish | Le*ff*e | 755858 | 34 | (Molinero et al., 2009) |
| Swedish | Saldo | 747959 | 38 | (Borin et al., 2008) |

Table 2: Information about the morphosyntactic lexicons used as external sources of lexical information in our MElt and MarMoT models. The number of entries and tagset sizes refers to the morphosyntactic lexicons we extracted and used in our models, not to the original resources.

Despite a few experiments published with MElt on languages other than French (Denis and Sagot, 2012; Le Roux et al., 2012; Seddah et al., 2013a), the original feature set used by MElt (standard and lexical features) was designed and tested mostly on this language, by building and evaluating tagging models on a variant of the French TreeBank. Since our goal was to carry out experiments in a multilingual setting, we have decided to design our own set of features, using the standard MElt features as a starting point. With respect to the original MElt feature set, we have added new ones, such as prefixes and suffixes of the following word, as well as a hybrid contextual feature obtained by concatenating the tag predicted for the preceding word and the tag(s) provided by the external lexicon for the following word.

In order to select the best performing feature set, we carried out a series of experiments using the multilingual dataset provided during the SPMRL parsing shared task (Seddah et al., 2013b). This included discarding useless or harmful features and selecting the maximal length of the prefixes and suffixes to be used as features, both for the current word and for the following word.[2]

We incorporated in MElt the best performing feature set, described in Table 1. All models discussed in this paper are based on this feature set.

# 3 Datasets

## 3.1 Corpora

We carried out our experiments on the Universal Dependencies v1.2 treebanks (Nivre and al., 2015), hereafter UD1.2, from which morphosyntactically annotated corpora can be trivially extracted. All UD1.2 corpora use a common tag set, the 17 *universal PoS tags*,[3] which is an extension of the tagset proposed by Petrov et al. (2012).

[2]During these tuning experiments, we used development sets for comparing feature sets, without evaluating any of our models on test sets.

[3]`http://universaldependencies.org/u/pos/all.html`

As our goal is to study the impact of lexical information for PoS tagging, we have restricted our experiments to UD1.2 corpora that cover languages for which we have morphosyntactic lexicons at our disposal, and for which Plank et al. (2016) provide results.[4] We considered UD1.2 corpora for the following 16 languages: Bulgarian, Croatian, Czech, Danish, English, French, German, Indonesian, Italian, Norwegian, Persian, Polish, Portuguese, Slovenian, Spanish and Swedish. Although this language list contains only one non-Indo-European (Indonesian), four major Indo-European sub-families are represented (Germanic, Romance, Slavic, Indo-Iranian). Overall, the 16 languages considered in our experiments are typologically, morphologically and syntactically fairly diverse.

### 3.2 Lexicons

We generate our external lexicons using the set of source lexicons listed in Table 2. Since external lexical information is exploited via features, there is no need for the external lexicons and the annotated corpora to use the same PoS inventory. Therefore, for each language, we simply extracted from the corresponding lexicon the PoS of each word based on its morphological tags, by removing all information provided except for its coarsest-level category.[5] We also added entries for punctuations when the source lexicons did not contain any.

We also performed experiments in which we retained the full original tags provided by the lexicons, with all morphological features included. On average, results were slightly better than those presented in the paper, although not statistically significantly. Moreover, the granularity of tag inventories in the lexicons is diverse, which makes it difficult to draw general conclusions about results based on full tags. This is why we only report results based on (coarse) PoS extracted from the original lexicons.

## 4 Experiments and results

### 4.1 Baseline models

In order to assess the respective contributions of external lexicons and word vector representations, we first compared the results of the three above-mentioned systems when trained without such additional lexical information. Table 3 provides the results of MElt and MarMoT retrained on UD1.2 corpora, together with the results publised on the same corpora by Plank et al. (2016), using their best model not enhanced by external word vector representations —i.e. the model they call $\vec{w} + \vec{c}$, which is a bidirectional LSTM that combines both word and character embeddings.

These results show that Plank et al.'s (2016) bi-LSTM performs extremely well, surpassed by MarMoT on only 3 out of 16 datasets (Czech, French and Italian), and by MElt only once (Indonesian).

### 4.2 Models enriched with external lexical information

Table 4 provides the results of four systems enriched with lexical information. The feature-based systems MElt and MarMoT, respectively based on MEMMs and CRFs, are extended with the lexical information provided by our morphosyntactic lexicons. This extension takes the form of additional features, as described in Section 2 for MElt. The results reported by Plank et al.

[4] They discarded all corpora containing fewer than 60k tokens in the training set, maybe as a result of the sensitivy of LSMTs to training set size.

[5] However, for French, we used the morphosyntactic variant of the Le*fff* that is included in the MElt distribution, and which relies on a variant of the French TreeBank known as FTB-UC (Candito and Crabbé, 2009).

| Model type<br>System | **MEMM**<br>**MElt** | **CRF**<br>**MarMoT** | **bi-LSTM**<br>**Plank et al.** |
|---|---|---|---|
| Bulgarian (bg) | 97.75 | 97.64 | **98.25** |
| Czech (cs) | 98.01 | **98.33** | 97.93 |
| Danish (da) | 95.48 | 95.56 | **95.94** |
| German (de) | 92.74 | 92.85 | **93.11** |
| English (en) | 94.06 | 94.37 | **94.61** |
| Spanish (es) | 95.32 | 95.14 | **95.34** |
| Persian (fa) | 96.72 | 96.43 | **96.89** |
| French (fr) | 95.81 | **96.13** | 96.04 |
| Croatian (hr) | 95.08 | 95.15 | **95.59** |
| Indonesian (id) | **93.74** | 93.63 | 92.79 |
| Italian (it) | 97.44 | **97.79** | 97.64 |
| Norwegian (no) | 96.68 | 97.26 | **97.77** |
| Polish (pl) | 96.12 | 96.21 | **96.62** |
| Portuguese (pt) | 97.38 | 97.43 | **97.48** |
| Slovene (sl) | 96.05 | 96.23 | **97.78** |
| Swedish (sv) | 95.97 | 96.03 | **96.30** |
| Macro-avg. | 95.90 | 96.01 | **96.26** |

Table 3: Overall accuracy (in %) of baseline systems, i.e. MElt and MarMoT models trained without external lexicons, and Plank et al.'s (2016) $\vec{c} + \vec{w}$ models, which do not make use of Polyglot embeddings. Best scores are highlighted for each corpus.

(2016) for their bidirectional LSTM when initialised with Polyglot embeddings trained on full wikipedias are also included, together with their new system FREQBIN, also initialised with Polyglot embeddings. FREQBIN trains bi-LSTMs to predict for each input word both a PoS and a label that represents its log frequency in the training data. As they word it, "the idea behind this model is to make the representation predictive for frequency, which encourages the model not to share representations between common and rare words, thus benefiting the handling of rare tokens."

The results, which are also displayed in Figures 1 and 2, show that all systems reach very similar results on average, although discrepancies can be observed from one dataset to another, on which we shall comment shortly. The best performing system in terms of macro-average is MElt (96.60%). Both bi-LSTM systems reach the same score (96.58%), the difference with MElt's results being non significant, whereas MarMoT is only 0.14% behind (96.46%). Given the better baseline scores of the neural approaches, these results show that the benefit of using external lexicons in the feature-based models MElt and MarMoT are much higher than those using Polyglot word vector representations as initialisations for bi-LSTMs.

Yet these very similar overall results reflect a different picture when focusing on OOV tagging accuracy. The best models for OOV tagging accuracy are, by far, FREQBIN models, which are beaten by MarMoT and by MElt only once each (on English and Danish respectively). The comparison on OOV tagging between MElt and MarMoT shows that MElt performs better on average than MarMoT, despite the fact that MarMoT's baseline results were better than those reached by MElt. This shows that the information provided by external morphosyntactic lexicons is better exploited by MElt's lexical features than by those used by MarMoT. On the other hand, the comparison of both bi-LSTM-based approaches confirm that the FREQBIN models is better by over 10% absolute on OOV tagging accuracy (94.28% vs. 83.59%), with 65% lower error rate.

One of the important differences between the lexical information provided by an external lexicon and word vectors built from raw corpora, apart from the very nature of the lexical

| Model type | **MEMM+lexicon** | | **CRF+lexicon** | | **bi-LSTM+Polyglot** | | **FREQBIN+Polyglot** | |
|---|---|---|---|---|---|---|---|---|
| System | **MElt** | | **MarMoT** | | **(Plank et al., 2016)** | | | |
| | overall | OOV | overall | OOV | overall | OOV | overall | OOV |
| Bulgarian (bg) | 98.15 | 93.95 | 98.05 | 93.06 | **98.23** | 87.40 | 97.97 | **97.37** |
| Czech (cs) | **98.58** | 94.83 | 98.48 | 93.68 | 98.02 | 89.02 | 97.89 | **94.91** |
| Danish (da) | 96.30 | **92.32** | 96.16 | 91.43 | 96.16 | 77.09 | **96.35** | 91.63 |
| German (de) | 93.43 | 88.08 | 93.10 | 87.21 | **93.51** | 81.95 | 93.38 | **90.97** |
| English (en) | 94.60 | 79.61 | 94.55 | **79.99** | **95.17** | 71.23 | 95.16 | 70.57 |
| Spanish (es) | 95.57 | 81.24 | 95.24 | 79.52 | 95.67 | 71.38 | **95.74** | **98.22** |
| Persian (fa) | 97.17 | 87.14 | 96.97 | 86.89 | **97.60** | 80.00 | 97.49 | **96.54** |
| French (fr) | 96.14 | 85.97 | **96.34** | 85.97 | 96.20 | 78.09 | 96.11 | **92.13** |
| Croatian (hr) | 96.70 | 93.01 | 96.19 | 91.23 | 96.27 | 84.62 | **96.82** | **97.29** |
| Indonesian (id) | **93.83** | 88.48 | 93.82 | 88.41 | 93.32 | 88.25 | 93.41 | **94.70** |
| Italian (it) | 97.82 | 91.98 | **98.03** | 91.82 | 97.90 | 83.59 | 97.95 | **98.46** |
| Norwegian (no) | 97.58 | 93.87 | 97.62 | 94.16 | **98.06** | 92.05 | 98.03 | **97.78** |
| Polish (pl) | **97.77** | 96.24 | 97.47 | 95.12 | 97.63 | 91.77 | 97.62 | **99.35** |
| Portuguese (pt) | 97.56 | 92.27 | 97.39 | 91.92 | **97.94** | 92.16 | 97.90 | **96.87** |
| Slovene (sl) | **97.53** | 96.50 | 97.23 | 94.89 | 96.97 | 80.48 | 96.84 | **95.63** |
| Swedish (sv) | **96.90** | 94.78 | 96.80 | 94.23 | 96.60 | 88.37 | 96.69 | **96.02** |
| Macro-avg. | **96.60** | 90.64 | 96.46 | 89.97 | 96.58 | 83.59 | 96.58 | **94.28** |

Table 4: Accuracy (in %) of the feature-based systems MElt and MarMoT as well as the two best LSTM-based systems by Plank et al. (2016) on UD1.2 datasets, which all use the 17 "universal PoS tags". MElt and MarMoT models integrate the external lexicons listed in Table 2, whereas bidirectional LSTM-based systems rely on Polyglot word embeddings. Best scores overall and on OOV words are highlighted for each corpus.

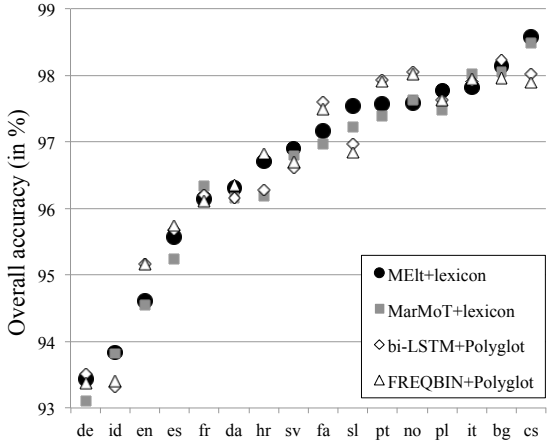



Figure 1: Graphical visualisation of the overall tagging accuracies for all four types of enriched models. Detailed results are given in Table 4. Languages are sorted by increasing MElt's overall tagging scores.

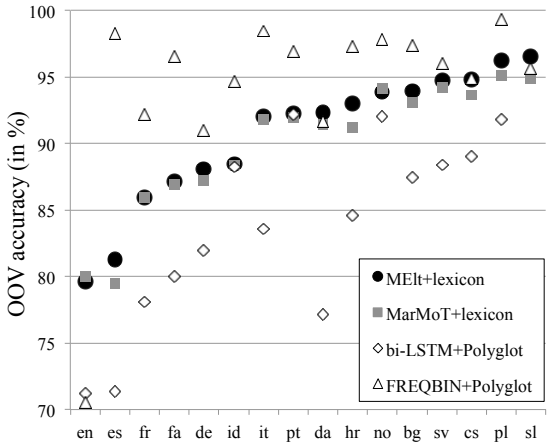



Figure 2: Graphical visualisation of the OOV tagging accuracies for all types of models enriched with external lexicons. Detailed results are given in Table 4. Languages are sorted by increasing MElt's OOV tagging scores.

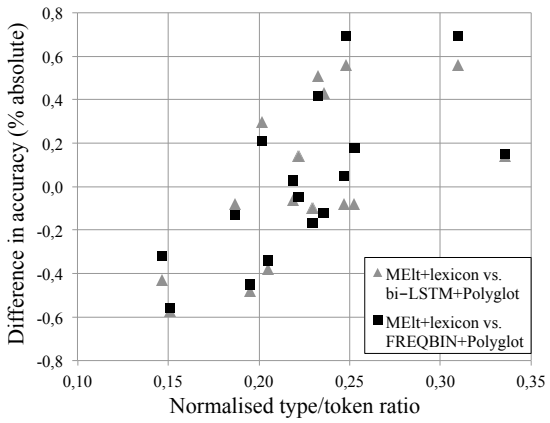



Figure 3: Difference between the tagging accuracy of lexicon-enhanced MElt models and each of the two types of Polyglot-enhanced neural bi-LSTM models plotted against training sets' normalised type/token ratio.

information provided, is the coverage and accuracy of this lexical information on rare words. All words in a morphosyntactic lexicon are associated with information of a same granularity and quality, which is not the case with word representations such as provided by Polyglot. Models that take advantage of external lexicons should therefore perform comparatively better on datasets containing a higher proportion of rarer words, provided the lexicons' coverage is high. In order to confirm this intuition, we have used a lexical richness metric based on the type/token ratio. Since this ratio is well-known for being sensitive to corpus length, we normalised it by computing it over the 60,000 first tokens of each training set. When this normalised type/token ratio is plotted against the difference between the results of MElt and both bi-LSTM-based models, the expected correlation is clearly visible (see Figure 3). This explains why MElt obtains better results on the morphologically richer Slavic datasets (average normalised type/token ratio: 0.28, average accuracy difference: 0.32 compared to both bi-LSTM+Polyglot and FREQBIN+Polyglot) and, at the other end of the spectrum, significantly worse results on the English dataset (normalised type/token ratio: 0.15, average accuracy difference: -0.56 compared to bi-LSTM+Polyglot, -0.57 compared to FREQBIN+Polyglot).

# 5 Conclusion

Two main conclusions can be drawn from our comparative results. First, feature-based tagging models adequately enriched with external morphosyntactic lexicons perform, on average, as well as bi-LSTMs enriched with word embeddings. Per-language results show that the best accuracy levels are reached by feature-based models, and in particular by our improved version of the MEMM-based system MElt, on datasets with high lexical variability (in short, for morphologically rich languages), whereas neural-based results perform better on datatsets with lower lexical variability (e.g. for English).

We have only compared the contribution of morphosyntactic lexicons to feature-based models (MEMMs, CRFs) and that of word vector representations to bi-LSTM-based models as reported by Plank et al. (2016). As mentioned above, work on the contribution of word vector representations to feature-based approaches has been carried out by Müller and Schütze (2015). However, the exploitation of existing morphosyntactic or morphological lexicons in neural models is a less studied question. Improvements over the state of the art might be achieved by integrating lexical information both from an external lexicon and from word vector representations into tagging models.

In that regard, further work will be required to understand which class of models perform the best. An option would be to integrate feature-based models such as a CRF with an LSTM-based layer, following recent proposals such as the one proposed by Lample et al. (2016) for named entity recognition.

# Contents

# References

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual nlp. In *Proc. of the Seventeenth Conf. on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria.

Ballesteros, M., Dyer, C., and Smith, N. A. (2015). Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(1):1137–1155.

Borin, L., Forsberg, M., and Lönngren, L. (2008). The hunting of the BLARK - SALDO, a freely available lexical database for swedish language technology. In *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, pages 21–32. Uppsala University, Uppsala, Sweden.

Braasch, A., Navarretta, C., Olsen, S., and Pedersen, B. S. (2008). Danish monolingual lexicon, documentation version 2.

Brants, T. (1996). Estimating markov model structures. In *Proc. of the 4th Conf. on Spoken Language Processing (ICSLP-96)*, pages 893–896.

Brants, T. (2000). TnT: A Statistical Part-of-speech Tagger. In *Proc. of the Sixth Conf. on Applied Natural Language Processing*, pages 224–231, Seattle, Washington, USA.

Candito, M. and Crabbé, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proc. of the 11th International Conf. on Parsing Technologies*, pages 138–141, Paris, France.

Chrupała, G. (2013). Text segmentation with character-level text embeddings. In *Proc. of the ICML Workshop on Deep Learning for Audio, Speech and Lang. Processing*, Atlanta, Georgia, USA.

Chrupała, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with morfette. In *Proc. of the 6th Language Resource and Evaluation Conf.*, Marrakech, Morocco.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of the 25th International Conf. on Machine Learning*, pages 160–167, Helsinki, Finland.

Constant, M. and Tellier, I. (2012). Evaluating the Impact of External Lexical Resources into a CRF-based Multiword Segmenter and Part-of-Speech Tagger. In *Proc. of LREC'12*, pages 646–650, Istanbul, Turkey.

Denis, P. and Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.

Erjavec, T. (2010). Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. of LREC 2010*, Valletta, Malta.

Goldberg, Y., Tsarfaty, R., Adler, M., and Elhadad, M. (2009). Enhancing unlexicalized parsing performance using a wide coverage lexicon, fuzzy tag-set mapping, and em-hmm-based lexical probabilities. In *Proc. of the 12th Conf. of the European Chapter of the ACL*, pages 327–335.

Hagen, K. and Nøklestad, A. (2010). Bruk av et norsk leksikon til tagging og andre språkteknologiske formål. *LexicoNordica*, 17:55–72.

Hajič, J. (2000). Morphological Tagging: Data vs. Dictionaries. In *Proc. of ANLP'00*, pages 94–101, Seattle, Washington, USA.

Hajič, J. and Hlaváčová, J. (2013). MorfFlex CZ. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neur. Comp.*, 9(8):1735–1780.

Kim, J.-D., Lee, S.-Z., and Rim, H.-C. (1999). HMM Specialization with Selective Lexicalization. In *Proc. of the join SIGDAT Conf. on Empirical Methods in Natural Lang. Processing and Very Large Corpora*.

Krek, S., Erjavec, T., and Holozan, P. (2008). Specifikacije za leksikon besednih oblik (kazalnik 3). Technical report, Projekt Sporazumevanje v slovenskem jeziku, Ljubljana, Slovenia.

Kübler, S., Scheutz, M., Baucom, E., and Israel, R. (2010). Adding context information to part of speech tagging for dialogues. In *Proc. of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 115–126.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

Lample, G., Ballesteros, M., Kawakami, K., Subramanian, S., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proc. of NAACL-HLT (NAACL 2016)*, San Diego, California, USA.

Le Roux, J., Sagot, B., and Seddah, D. (2012). Statistical parsing of spanish and data driven lemmatization. In *Proc. of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*, Jeju, Korea.

Ling, W., Luís, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., Black, A. W., and Trancoso, I. (2015). Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proc. of the 2015 Conf. on Empirical Methods in Natural Lang. Processing*, pages 1520–1530, Lisbon, Portugal.

Magerman, D. M. (1995). Statistical decision-tree models for parsing. In *Proc. of the 33rd Annual Meeting on ACL*, pages 276–283, Cambridge, Mass., USA.

Merialdo, B. (1994). Tagging english text with a probabilistic model. *Computational Linguistics*, 20(2):155–171.

Molinero, M. A., Sagot, B., and Nicolas, L. (2009). A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe. In *Proc. of the 7th conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA.

Müller, T. and Schütze, H. (2015). Robust morphological tagging with word representations. In *Proc. of the 2015 Conf. of the North American Chapter of the ACL: Human Language Technologies*, Denver, Colorado, USA.

Nivre, J. and al. (2015). Universal dependencies 1.2. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Oliver, A. and Tadić, M. (2004). Enlarging the Croatian morphological lexicon by automatic lexical acquisition from raw corpora. In *Proc. of LREC 2004*, pages 1259–1262, Lisbon, Portugal.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proc. of LREC 2012*, Istanbul, Turkey.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proc. of the 54th Annual Meeting of the ACL*, Berlin, Germany. To appear.

Ranchhod, E., Mota, C., and Baptista, J. (1999). A Computational Lexicon of Portuguese for Automatic Text Parsing. In *Proc. of the SIGLEX99 workshop on Standardizing Lexical Resources*, College Park, Maryland, USA.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proc. of International Conf. on Empirical Methods in Natural Language Processing*, pages 133–142.

Sagot, B. (2007). Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proc. of LTC 2005*, pages 423–427, Poznań, Poland.

Sagot, B. (2010). The Le*fff*, a freely available, accurate and large-coverage lexicon for french. In *Proc. of LREC 2010*, Valletta, Malta.

Sagot, B. (2014). DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German. In *Language Resources and Evaluation Conf.*, Reykjavik, Iceland.

Sagot, B. and Walther, G. (2010). A morphological lexicon for the Persian language. In *Proc. of LREC 2010*, Valletta, Malta.

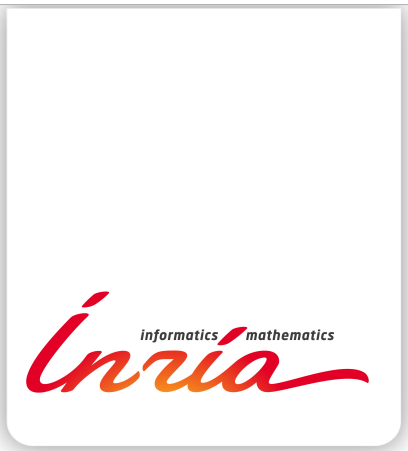
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc. of International Conf. on New Methods in Language Processing*, Manchester, UK.

Seddah, D., Le Roux, J., and Sagot, B. (2013a). Data driven lemmatization and parsing of italian. In Magnini, B., Cutugno, F., Falcone, M., and Pianta, E., editors, *Evaluation of Natural Language and Speech Tools for Italian (Revised Selected Papers)*, volume 7689 of *Lecture Notes in Computer Science*, pages 249–256. Springer-Verlag, Rome, Italy.

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., and Villemonte de La Clergerie, E. (2013b). 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.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. of International Conf. on New Methods in Language Processing*, pages 63–70, Hong Kong, China.

Zanchetta, E. and Baroni, M. (2005). Morph-it! a free corpus-based morphological resource for the Italian language. In *Proc. of the Corpus linguistics Conf.*, pages 1–12, Birmingham, UK.




**RESEARCH CENTRE**
**PARIS – ROCQUENCOURT**

Domaine de Voluceau, - Rocquencourt
B.P. 105 - 78153 Le Chesnay Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399