A multilingual collection of CoNLL-U-compatible morphological lexicons

Benoît Sagot

A multilingual collection of CoNLL-U-compatible morphological lexicons

Benoît Sagot
Inria
2 rue Simone Iff, CS 42112, 75589 Paris Cedex 12, France
benoit.sagot@inria.fr

Abstract

We introduce UDLexicons, a multilingual collection of morphological lexicons that follow the guidelines and format of the *Universal Dependencies* initiative. We describe the three approaches we use to create 53 morphological lexicons covering 38 languages, based on existing resources. These lexicons, which are freely available, have already proven useful for improving part-of-speech tagging accuracy in state-of-the-art architectures.

Keywords: Morphological Lexicons, Universal Dependencies, Freely Available Language Resources

1. Introduction

Morphological information belongs to the most fundamental types of linguistic knowledge. It is often either encoded into morphological analysers or gathered in the form of morphological lexicons. Such lexicons, which constitute the focus of this paper, are collections of lexical entries that typically associate a wordform with a part-of-speech (or morphosyntactic category), morphological features (such as gender, tense, etc.) and a lemma. Beyond direct lexicon lookup, used in virtually all types of natural language processing applications and computational linguistic studies, morphological lexicons have been shown to significantly improve tasks such as part-of-speech tagging and parsing.

There is currently no universally accepted way to encode morphological lexical information. Past multilingual projects such as MULTEXT/MULTEXT-East (Ide and Véronis, 1994; Erjavec, 2010) have resulted in the publication of morphological lexicons for a number of languages based on the same set of categories and morphosyntactic features, but they are still limited in scope.

Yet another type of language resource embeds morphological lexical information, namely treebanks. This type of resource has recently seen the emergence of a *de facto* trans-lingual standard and the publication of an increasing number of treebanks for numerous languages following a universal set of guidelines, encoded in the CoNLL-U format and gathered under the name *Universal Dependencies* (hereafter UD).[1] This treebank collection (Nivre et al., 2016; Nivre et al., 2017) follows several previous initiatives, such as the proposal of a universal part-of-speech tagset (Petrov et al., 2012) and the multilingual datasets released in the context of several shared tasks and projects (Buchholz and Marsi, 2006; Nivre et al., 2007; Zeman et al., 2012; Seddah et al., 2013).

The UD initiative has therefore allowed a simpler, unified access to treebank resources, giving a new impetus to research in topics such as multilingual and cross-lingual tokenisation, part-of-speech tagging, dependency parsing and quantitative linguistics. It is therefore important for morphological lexicons, another major source of linguistic information for such tasks, to also be available for many languages following a universal set of guidelines. The obvious choice would be to make use of the UD guidelines themselves.

We have therefore developed a multilingual collection of morphological lexicons that follow the UD guidelines regarding part-of-speech and morphological features. We used three main sources of lexical information:

- In the context of the CoNLL 2017 UD morphological and syntactic analysis shared task (Zeman et al., 2017) based on UD treebank data, we used lexical information available in the Apertium[2] (Forcada et al., 2011) and Giellatekno[3] projects. This information, which consists of morphological lexicons and analysers, allowed us to provide additional features to our part-of-speech tagging architecture, with high-accuracy results (Villemonte de La Clergerie et al., 2017). In some cases, information from Apertium or Giellatekno converted into the UD format was complemented with information extracted from the training sections of the UD treebanks (v2.0).
  We also developed a simple, unsupervised yet original algorithm for transferring morphological lexical information from a resourced language to a closely-related one not covered by Apertium and Giellatekno.

- We converted into the UD format a variety of freely available lexicons, among which the lexicons developed in the Alexina framework (Sagot, 2010).[4]

- We used the UD treebanks themselves (v2.0) in order to complete our lexicons with two types of information: (i) multi-word tokens and (ii) entries for categories (UPOS) not covered by the lexical sources mentioned above.

Our contribution thereby lies in a collection of 53 UD-compatible lexicons covering 38 distinct languages and the

[1] http://www.universaldependencies.org

[2] https://svn.code.sf.net/p/apertium/svn/languages

[3] https://victorio.uit.no/langtech/trunk/langs

[4] These resources were not allowed in the CoNLL 2017 UD parsing shared task and were not used in this context.

methods used to create this collection.[5] They are encoded in the CoNLL-UL format, an extension of the CoNLL-U format introduced by More et al. (2018) aimed at representing morphological information, in particular the output of tokenisation and morphological analysis tools, even when they are non-deterministic.

In the remainder of this paper, we first describe how we converted Apertium and Giellatekno morphological analysers and lexicons into CoNLL-UL morphological lexicons. We also briefly sketch our cross-lingual transfer algorithm aimed at building lexicons for poorly resourced languages. We then briefly explain how we converted other existing morphological lexicons to CoNLL-UL. Next we provide a summary of the lexicons we obtained and the languages covered. In the conclusion, we mention a few results on the use of some of these lexicons in part-of-speech tagging experiments, using both statistical and neural approaches.

2. CoNLL-UL lexicon creation

2.1. Creation of CoNLL-UL lexicons from Apertium and Giellatekno resources

The Apertium and Giellatekno projects publish freely available tools and resources for a growing number of languages. These projects rely on morphological analysers, which, in turn, are primarily based either on morphological lexicons or on finite-state morphological grammars. Each language makes use of its own guidelines regarding the inventory of categories and the detailed definition of morphological features and feature values. Yet many categories, features and feature values are used consistently across languages. This allowed us to manually develop a single conversion script that interprets Apertium and Giellatekno categories and morphological features and rewrites them in terms of the UD guidelines, using the Universal Part-Of-Speech (hereafter UPOS)[6] and the Universal morphological Feature (UFEAT)[7] inventories.

Before this conversion could take place, morphological lexical information had to be extracted from the Apertium or Giellatekno resources. Depending on the type of resource available for a given language, we adopted one of the two following strategies:

- Direct extraction and reformatting of the monolingual morphological lexicon provided by Apertium (lexicon type code "AP" in Table 2);
- Automatic morphological analysis of the raw monolingual corpora provided by the ConLL 2017 shared task organisers, using Apertium or Giellatekno morphological analysers (codes "APma" or "GTma"). More precisely, we downloaded the corresponding monolingual part of OPUS's OpenSubtitles2016 corpus,[8] tokenised it using a basic language-independent rule-based tokeniser, extracted the 1 million most frequent tokens, and retrieved all their morphological analyses by the corresponding morphological analyser provided by Apertium (or, failing that, Giellatekno). All these analyses were then gathered in the form of a lexicon.

We applied the direct conversion technique to 27 lexicons/languages, whereas the approach based on morphological analysers allowed us to cover 4 additional languages (2 via Apertium, 2 via Giellatekno). We also experimented the extension of these lexicons with data from the shared task training sets and the raw corpus automatically analysed by the UDPipe tool (Straka et al., 2016) by the organisers. Whenever it helped increasing our tagging results, as measured on the development sets, we applied these extensions. They are respectively indicated in Table 2 by "+T" and "+U" (this also applies to lexicons created in the next section).

For a few languages, we also created expanded versions of our lexicons using word embeddings re-computed on the raw data provided by the CoNLL 2017 shared task organisers. We assigned to words unknown to the lexicon the morphological information associated with the closest known word (using a simple Euclidian distance in the word embedding space).[9] When the best performing lexicon is one of these extended lexicons, it is indicated in Table 2 by the "-e" suffix.

2.2. Unsupervised cross-lingual transfer of Apertium and Giellatekno lexicons

In the context of our participation to the CoNLL 2017 shared task (Villemonte de La Clergerie et al., 2017), we were interested in producing as many morphological lexicons as possible for all languages involved, in order to optimise part-of-speech tagging and morphological annotation as much as possible. Yet the closed setting disallowed the use of any resources outside a pre-defined list, which included Apertium and Giellatekno monolingual morphological resources. For several languages, including Slovak, these projects did not provide adequate resources. However, Apertium does include a morphological lexicon for Czech, a language closely related to Slovak. We therefore decided to set up an unsupervised cross-lingual transfer technique to produce a Slovak morphological lexicon from the Apertium Czech one. In doing this, we rely on the hypothesis that the same morphological features and feature values are valid for Slovak as for Czech, and that part-of-speech and morphological features are stable across word-alignment links. This technique is therefore more relevant for closely related languages that are typologically similar.

The main cross-lingual resource we were allowed to use was the OpenSubtitles2016 corpus set, which provides monolingual as well as sentence-aligned bilingual parallel subtitle corpora. We took advantage of the parallel

[5] All resulting lexicons are available as free resources via the following website, together with their respective licences: http://pauillac.inria.fr/ sagot/udlexicons.html.

[6] http://universaldependencies.org/u/pos/

[7] http://universaldependencies.org/u/feat/

[8] Exploiting this dataset was allowed as per the CoNLL 2017 shared task rules.

[9] We did not used the embeddings provided by the organisers because we experimentally found that the 10-token window used to train these embeddings resulted in less accurate results than when using smaller windows, especially when the raw corpus available was of a limited size.

data in two steps, which we now illustrate on the example of Slovak and Czech. Firstly, we performed an endogenous, unsupervised extraction of a bilingual lexicon from the Slovak-Czech OpenSubtitles2016 parallel corpus, after a basic tokenisation using the same tool as mentioned above. For this extraction process, we defined a matching metric between Czech tokens and Slovak tokens, which takes into account both the distribution of the tokens across sentences—the more often the Slovak token is found in sentences aligned with Czech sentences containing the Czech token, the higher the score—and the weighted Levenshtein edit distance between the two tokens, in order to give an advantage to cognate pairs.[10]

We first computed this distance using default weights for the Levenshtein distance (all operations cost 1). Character-to-character alignments produced by the Levenshtein algorithm allowed us to update the weight matrix[11] and recomputed all matching metrics. After reaching stability (i.e. after a few iterations) and metric-based thresholding, we obtained a bilingual Czech-Slovak lexicon. We then retrieved part-of-speech and morphological features for each Czech word in our bilingual lexicon and projected it onto all Slovak words it is aligned with.

We applied this technique to the following language pairs, in which the source (resourced) language is indicated first: Czech-Slovak, Italian-Latin, Russian-Ukrainian, Slovene-Croatian.The evaluation of the accuracy of this transfer technique is currently ongoing for the Czech-Slovak pair. Task-based evaluation of the output lexicons by using them as additional sources of information for a statistical part-of-speech tagger has already proven successful (see Section 3.). In Table 2 we use the code "TR*source_language*" to identify lexicons created in this cross-language way.

[10]More formally, we define the distributional distance $DS(s,t)$ between a token s in the source (resourced) language and a token t in the target language as

$$DS(s,t) = 2 \cdot \left(\frac{\mathrm{occ}(s)}{\mathrm{nbsentpairs}(s,t)} + \frac{\mathrm{occ}(t)}{\mathrm{nbsentpairs}(s,t)} \right)^{-1},$$

where $\mathrm{nbsentpairs}(s,t)$ is the number of sentence pairs (a source sentence aligned with a target sentence) such that the source sentence contains s and the target sentence contains t. This distance equals 1 if and only if all occurrences of s are in sentences aligned with sentences containing t, and vice versa. Next, we call $d_{\vec{w}}(s,t)$ the weighted Levenshtein distance between s and t, where $\vec{w}$ stores the weight of each possible operation (e.g. replacing "ŕ" by "r"). Finally, our matching metric is defined as $M(s,t) = DS(s,t) - \frac{1}{10} \cdot d^2_{\vec{w}} \cdot \max\left(\frac{1}{10}, \frac{1}{2} - DS\right)$. The idea underlying this metric is that s and t are likely to be translations of each another if their distributional score is high or if they are formally close, this latter criterion becoming more important if the distributional score decreases.

[11]The weight of an operation tranforming a source character c_s (or the empty string in the case of an insertion) into a target character c_t (or the empty string in the case of a deletion) is defined as $1 - \sqrt{\frac{\mathrm{transfocc}(c_s,c_t)}{\mathrm{occ}(c_s)}}$, where $\mathrm{transfocc}(c_s,c_t)$ is the number of times c_s was transformed into c_t in the previous iteration, and $\mathrm{occ}(c_s)$ is the total number of c_s's in the source lexicon.

2.3. Creation of CoNLL-UL lexicons from other freely available lexicons

Independently of the Apertium and Giellatekno projects, many research teams have developed large-scale morphological lexicons for a variety of languages. We therefore designed a conversion process aimed at producing high-quality CoNLL-UL lexicons from these existing resources. The process is the following:

1. For each word, we register its corresponding ⟨category, morphological feature values, lemma⟩ triplets found in the source lexicon. We also extract its corresponding ⟨UPOS, lemma⟩ pairs from the UD v2 training set from the language's main treebank.[12] We then extract the most frequent UPOS for each source ⟨category, morphological feature values⟩ pair, based on identical ⟨wordform, lemma⟩ pairs.

2. We then apply heuristics based on the source-lexicon-to-UPOS correspondence obtained in the first step to overcome lemmatisation mismatches between the source lexicon and the UD treebank. We store the corresponding ⟨wordform, category, morphological features, source lemma, UD lemma⟩ 5-tuples for later use. We also update the ⟨source category, source morphological feature values, UPOS⟩ triples created during step 1 with the results of this step.

3. We use these updated ⟨source category, source morphological feature values, UPOS⟩ as a basis for extracting full correspondence patterns, i.e. 4-tuples of the form ⟨source category, source morphological feature values, UPOS, UFEAT⟩. More precisely, for each triple of the form ⟨source category, source morphological feature values, UPOS⟩ we select the most frequent UFEAT in the UD data among those associated with wordforms known to the lexicon with the category and morphological feature values at hand.

4. We output the updated ⟨source category, source morphological values, UPOS, UFEAT⟩ 4-tuples in an Excel file for manual reviewing and completion. We also include ⟨source category, source morphological values⟩ pairs for which no UPOS was found—often because they are not attested in the UD data—as well as UPOS for which no such triplet was created. This allows for the manual work to be exhaustive. We then perform a full manual review and correction of the Excel correspondence file.

5. We use the manually corrected file to automatically convert the source lexicon into a CoNLL-UL lexicon.

We applied this strategy on 18 freely available lexicons (see Table 2):

- 8 lexicons developed in the Alexina framework (Sagot, 2010), covering French (Sagot, 2010, Lefff), Polish (Sagot, 2007, PolLex), Slovak (Sagot, 2005,

[12]I.e. the treebank whose identifier is the language code itself (e.g. `fr` rather than `fr-sequoia`).

From	To	Form or Token	Lemma	UPOS	CPOS	UFEAT	Misc
0	1	*encodent*	encoder	VERB	_	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin	_
0-2		auxquels					
0	1	*à*	à	ADP	_	_	_
1	2	*lesquels*	lequel	PRON	_	Gender=Masc|Number=Plur	_

Table 1: Two entries resulting from the conversion of the Lefff into the CoNLL-UL format (for space reasons, the CPOS column is displayed as if it were empty).

SkLex), Spanish (Molinero et al., 2009, Leffe), Galician (Leffga), Persian (Sagot and Walther, 2010, PerLex) German (Sagot, 2014, DeLex) and English (EnLex);

- 10 other lexicons covering Italian (Zanchetta and Baroni, 2005, Morph-it!), Swedish (Borin et al., 2008, saldo), Ancient Greek (Heslin, 2007, Diogenes Ancient Greek lexicon), Latin (Heslin, 2007, Diogenes Latin lexicon), Croatian (Oliver and Tadić, 2004, hml), Irish (Měchura, 2014, INMDB), Norwegian (Bokmål) (The Language Council of Norway, 2011, OrdBank-BM), Portuguese (Ranchhod et al., 1999, Labellex-PT) and Slovenian (Krek et al., 2008, SloLeks).

Some of the above-listed Alexina lexicons include information about multi-word tokens. For instance, the French token auxquels 'to which$_{PL}$' is described in Lefff as the contraction of the two wordforms *à* 'to' and *lesquels* 'which$_{PL}$'. Moreover, the part-of-speech of the wordforms involved is sometimes specified. We automatically extended our converted lexicons with CoNLL-UL entries for these multi-word tokens by combining existing entries for the underlying wordforms. Whenever part-of-speech information is provided for a wordfrom, we limit ourselves to entries with the corresponding UPOS. The entry generated for auxquels is shown in Table 1.

2.4. Final treebank-based extension

We complemented all lexicons created using one of the three techniques described above with information extracted from the training part of the UD treebanks. For each language, we first extracted all entries, both simple entries and multi-word tokens, from the training part of the corresponding UD treebank.[13]

Next, in order to limit the bias towards UD training data, we automatically computed a frequency threshold and discarded those entries occurring less frequently.[14]

Finally, we discard all simple entries whose UPOS is already attested in our lexicon. The reason for this is that we want to preserve the consistency of our lexicons in terms of morphological annotation guidelines (UFEAT). UFEAT annotations in UD treebanks is not always consistent with UFEAT information found in our lexicons (for instance, a feature can be provided in one resource and not in the other). Merging entries extracted from UD treebanks with entries already present in our lexicons could therefore result in inconsistencies. Doing so for UPOS not yet covered (e.g. PUNCT in several Apertium- and Giellatekno-based lexicons) does not create such inconsistencies.

2.5. Results

Table 2 gives quantitative information about the lexicons produced by the three methods described in Sections 2.1. to 2.3. followed by the extension step described in Section 2.4.. They constitute the version 0.2 of the UDLexicons collection.

3. Preliminary task-based evaluation in part-of-speech taggers

The CoNLL-UL lexicons we extracted from Apertium and Giellatekno data, including those created via cross-lingual transfer, served as a source of additional features in two different part-of-speech tagging experiments. The first one is our participation in the CoNLL 2017 shared task. For this shared task, we wanted to explore many ways of creating such additional features from lexicons, and to compare different lexicon variants (cf. the "+U" and "+T" extensions). We therefore developed a new statistical MEMM tagger, following our previous work in this direction (Denis and Sagot, 2012), but this time based on the Vowpal Wabbit architecture. We selected our pre-processing architecture based on parsing results on the development sets. In particular, we chose to use either this new tagger or the UDPipe baseline provided by the organisers. Although our new tagger had higher accuracies than UDPipe on all datasets but 4 (Villemonte de La Clergerie et al., 2017, Table 1), we ended up using our tagger on only half of the testing datasets. This allowed us to be ranked 3/33 for UPOS tagging.[15,16]

[13] For several languages, more than one UD treebank is available. In such cases, we use the concatenatation of the training parts of all of them.

[14] Our threshold is computed using the following heuristics: we order entries in decreasing order of frequency, and identify the minimum number of entries necessary to cover 90% of the data, starting from the most frequent one and accumulating entries in decreasing order of frequency. We then extract the number of occurrences $\text{occ}_{90\%}$ of the last selected entry. If $\text{occ}_{90\%} = 1$, which typically occurs on small datasets, we also compute $\text{occ}_{75\%}$. We then fix our frequency threshold via a minimum number of occurrences $\text{occ}_{\text{threshold}}$ defined as follows:

$$\text{occ}_{\text{threshold}} = \begin{cases} \max(3, \text{occ}_{90\%}) & \text{if } \text{occ}_{90\%} > 1 \\ \max(2, \text{occ}_{75\%}) & \text{if } \text{occ}_{90\%} = 1 \end{cases}$$

[15] More recent, unofficial results using improved parsers have resulted in our tagger being selected more often rather than UDPipe, thus further improving our overall UPOS tagging results.

[16] Our UFEAT scores in this shared task are not meaningful, because we explicitly decided to only predict a subset of all morphological features.

Apertium/Giellatekno-based resources				
lang.	**type**	**#simple entries**	**#complex entries**	**#distinct wforms**
ar	AP-e	660K	3,055	246K
bg	AP	93K		76K
ca	AP-e_{100}	381K	47	261K
cs	AP	1,875K	10	480K
da	AP	683K		377K
de	AP	2,180K	28	411K
el	AP	47K		29K
en	AP	127K	1	96K
es	AP	325K	161	273K
et	GTma	44K		33K
eu	AP	49K		45K
fi	GTma	228K	17	156K
fr	AP-e^6_{1000}	156K	21	123K
gl	AP	241K	28	191K
he	AP	268K	4425	206K
hi	AP	159K		66K
hr	TRsl	14K		11K
id	AP	12K		12K
it	AP	278K	105	229K
kk	APma	434K		274K
la	TRit+T-e_{100}	13K		11K
lv	AP	314K		166K
nl	AP	167K		78K
no	AP	2,470K		1,373K
pl	AP	1,316K	291	525K
pt	AP	159K	184	119K
ro	AP	229K		151K
ru	AP	4,401K		2,159K
sk	TRcs	66K		36K
sl	AP	654K		203K
sv	AP	2,378K		1,319K
tr	APma	417K	697	246K
uk	TRru	23K		12K
ur	AP	98K		54K
zh	AP+U	17K		10K

Other resources				
lang.	**type**	**#simple entries**	**#complex entries**	**#distinct wforms**
de	DeLex	1,138K	282	264K
en	EnLex	695K	186	508K
es	Le*ff*e	843K	163	680K
fa	PerLex	178K	68	168K
fr	Le*fff*	650K	22	456K
ga	INMDB	57K		39K
gl	Le*ff*ga	949K	28	402K
grc	Diogenes	4,490K		1,004K
hr	HML5	3,854K		1,208K
it	Morph-it!	785K	105	378K
la	Diogenes	1,870K		425K
nl	Alpino	122K		73K
no	OrdBank$_{\mathrm{BM}}$	878K		636K
pl	PolLex	1,368K	291	355K
pt	labellex-pt	1,841K	184	820K
sk	SkLex	750K		419K
sl	SloLeks	2,626K		880K
sv	Saldo	1,241K		701K

Table 2: CoNLL-UL lexicons, version 0.2 (see text for an explanation of the 'type' column and for references).

In another recent experiment (Sagot and Martínez Alonso, 2017), we have shown that these lexicons also improve tagging results when a state-of-the-art neural architecture, namely that of Plank et al. (2016), is extended to take into account external lexical information, even when word encodings and character-based encodings are used.

4. Conclusion and future steps

We have developed a collection of morphological lexicons compatible with the Universal Dependencies guidelines and format. These lexicons are freely available, under licences that depend on those of the original resources. Such a lexicon collection could serve as a starting point for providing the community with a set of lexical resources that will consistently complement the UD treebank collection, as well as the morphological analysers developed as companions to the CoNLL-UL proposal (More et al., 2018). These initiatives give easier access to morphological information in multilingual settings in contexts such as parsing and information extraction.

Our future steps are now threefold. Firstly, as mentioned above, the conversion mappings from the original categories and features to UPOS and UFEATs must be thoroughly reviewed and improved, in a way consistent with the UD treebanks. Secondly, the converted lexicons will be carefully evaluated, both in terms of precision and coverage. Thirdly, the availability of different lexicons for the same languages will make it possible to merge them in different ways, in order to optimise their coverage and accuracy. This is particularly true for lexicons that are likely to be less reliable, such as those created using the cross-lingual transfer technique described in Section 2.2. Finally, more lexicons will be included in the collection over time, such as the MULTEXT/MULTEXT-East lexicons that are distributed under a free licence.

5. References

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.

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proc. of the Tenth Conference on Computational Natural Language Learning*, pages 149–164, New York City, USA.

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 *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC'2010)*, Valletta, Malta.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Heslin, P. J. (2007). Diogenes, version 3.1. http://www.dur.ac.uk/p.j.heslin/Software/Diogenes/.

Ide, N. and Véronis, J. (1994). MULTEXT: Multilingual Text Tools and Corpora. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan.

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

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

More, A., Çetinoğlu, O., Çöltekin, c., Habash, N., Sagot, B., Seddah, D., Taji, D., and Tsarfaty, R. (2018). CoNLL-UL: Universal Morphological Lattices for Universal Dependency Parsing. In *Proc. of LREC 2018*, Miyazaki, Japan.

Měchura, M. B. (2014). Irish National Morphology Database: A High-Accuracy Open-Source Dataset of Irish Words. In *Proc. of the Celtic Language Technology Workshop at CoLing*, Dublin, Ireland.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bosco, C., Bouma, G., Bowman, S., Candito, M., Cebiroğlu Eryiğit, G., Celano, G. G. A., Chalub, F., Choi, J., Çöltekin, Ç., Connor, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Erjavec, T., Farkas, R., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hà Mỹ, L., Haug, D., Hladká, B., Hohle, P., Ion, R., Irimia, E., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kotsyba, N., Krek, S., Laippala, V., Lê Hồng, P., Lenci, A., Ljubešić, N., Lyashevskaya, O., Lynn, T., Makazhanov, A., Manning, C., Mărănduc, C., Mareček, D., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., More, A., Mori, S., Moskalevskyi, B., Muischnek, K., Mustafina, N., Müürisep, K., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nurmi, H., Ojala, S., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Real, L., Rituma, L., Rosa, R., Saleh, S., Sanguinetti, M., Saulīte, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shakurova, L., Shen, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Tsarfaty, R., Tyers, F., Uematsu, S., Uria, L., van Noord, G., Varga, V., Vincze, V., Washington, J. N., Žabokrtský, Z., Zeldes, A., Zeman, D., and Zhu, H. (2017). Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

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 Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

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

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.

Sagot, B. and Martínez Alonso, H. (2017). Improving neural tagging with lexical information. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 25–31, Pisa, Italy.

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

Sagot, B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658, Proc. of TSD'05*, pages 156–163, Karlovy Vary, République tchèque. Springer-Verlag.

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 L*efff*, a freely available, accurate and large-coverage lexicon for french. In *Proceedings of the 7th international conference on Language Resources and Evaluation (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.

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. (2013). Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

The Language Council of Norway. (2011). Norsk Ordbank in Norwegian Bokmål. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository, http://hdl.handle.net/11509/136.

Villemonte de La Clergerie, É., Sagot, B., and Seddah, D. (2017). The ParisNLP entry at the ConLL UD Shared Task 2017: A Tale of a #ParsingTragedy. In *Conference on Computational Natural Language Learning*, Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 243–252, Vancouver, Canada, August.

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.

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). Hamledt: To parse or not to parse? In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., and Li, J. (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 1–19, Vancouver, Canada, August. Association for Computational Linguistics.