Neural Lemmatization of Multiword Expressions

Marine Schmitt
Université de Lorraine & CNRS
ATILF
F-54000 Nancy, France
Marine.Schmitt@atilf.fr

Mathieu Constant
Université de Lorraine & CNRS
ATILF
F-54000 Nancy, France
Mathieu.Constant@univ-lorraine.fr

Abstract

This article focuses on the lemmatization of multiword expressions (MWEs). We propose a deep encoder-decoder architecture generating for every MWE word its corresponding part in the lemma, based on the internal context of the MWE. The encoder relies on recurrent networks based on (1) the character sequence of the individual words to capture their morphological properties, and (2) the word sequence of the MWE to capture lexical and syntactic properties. The decoder in charge of generating the corresponding part of the lemma for each word of the MWE is based on a classical character-level attention-based recurrent model. Our model is evaluated for Italian, French, Polish and Portuguese and shows good performances except for Polish.

1 Introduction

Lemmatization consists in finding the canonical form of an inflected form occurring in a text. Usually, the lemma is the base form that can be found in a dictionary. In this paper, we are interested in the lemmatization of multiword expressions (MWEs), that has received little attention in the past. MWEs consist of combinations of several words that show some idiosyncrasy (Gross, 1986; Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017). They display the linguistic properties of a lexical unit and are present in lexicons as simple words are. For instance, such a task may be of interest for the identification of concepts and entities in morphologically-rich languages.[1]

The main difficulty of the task resides in the variable morphological, lexical and syntactic properties of MWEs leading to many different lemmatization rules on top of simple-word lemmatization knowledge, as illustrated by the 27 hand-crafted rules used by the rule-based multiword lemmatizer for Polish described in Marcińczuk (2017). For example, in French, the nominal MWE *cartes bleues* (cards.noun.fem.pl blue.noun.fem.pl), meaning *credit cards*, is lemmatized in *carte bleue* (car.noun.fem.sg blue.adj.fem.sg) where the adjective *bleue* (blue) agrees in person (sg) and gender (fem) with the noun *carte* (card). A single-word lemmatization would not preserve the gender agreement in this example: the feminine adjective *bleues* would be lemmatized in the masculine *bleu*.

In this paper, we propose a deep encoder-decoder architecture generating for every MWE word its corresponding part in the lemma, based on the internal context of the MWE. The encoder relies on recurrent networks based on (1) the character sequence of the individual words to capture their morphological properties, and (2) the word sequence of the MWE to capture lexical and syntactic properties. The decoder in charge of generating the corresponding part of the lemma for each word of the MWE is based on a classical character-level attention-based recurrent model. One research question is whether the system is able to encode the complex linguistic properties in order to generate an accurate MWE lemma. As a preliminary stage, we evaluated our architecture in five suffix-based inflectional languages with a special focus on French and Polish.

Contrary to the lemmatization of simple words (Bergmanis and Goldwater, 2018), our task is not a disambiguation task[2], as for a given MWE form, there is one possible lemma in all cases but some very rare exceptions. This means that the lemma

[1] Different shared tasks including lemmatization for Slavic languages have been organized recently: PolEval 2019 shared task on lemmatization of proper names and multi-word phrases, BSNLP 2019 shared task on multilingual named entity recognition including lemmatization.

[2] Note that MWE lemmatization requires, as previous step, MWE identification which involves disambiguation.

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 142–148
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

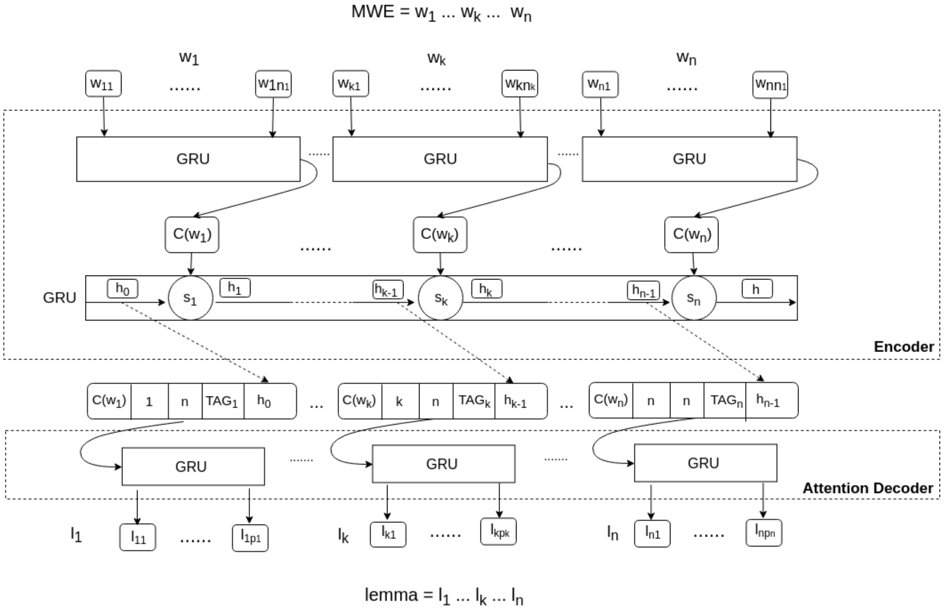

Figure 1: Neural architecture. For simplification, we do not show hidden and softmax layers of attention decoder. We use ReLU as activation of the hidden layer. TAG$_k$ stands for the embedding of the predicted POS tag of the word w_k, possibly concatenated with the embedding of the gold MWE-level POS tag.

of a known MWE is simply its associated lemma in the training data. The interest of a neural system is thus limited to the case of unknown MWEs. One research question is whether the system is able to generalize well on unknown MWEs.

To the best of our knowledge, this is the first attempt to implement a language-independent MWE lemmatizer based entirely on neural networks. Previous work used rule-based methods and/or statistical classification methods (Piskorski et al., 2007; Radziszewski, 2013; Stankovic et al., 2016; Marcińczuk, 2017).

The article is organized as follows. First, we describe our model and our dataset. Then we display and discuss experimental results, before describing related work.

2 Model

Our lemmatization model is based on a deep encoder-decoder architecture as shown in Figure 1. The input MWE is a sequence w_1 w_2 ... w_n of n words. It is given without any external context as there is no disambiguation to perform (cf. section 1). Every word w_k is decomposed in a sequence w_{k1} w_{k2} ... w_{kn_k} of n_k characters that is passed to a Gated Recurrent Unit[3] (GRU) that outputs a character-based word embedding $C(w_k)$, which corresponds to the output of the last GRU cell. The whole MWE sequence $C(w_1)$ $C(w_2)$ ... $C(w_n)$ is then passed to a GRU[4] in order to capture the internal context of the MWE. For every word w_k, a decoder generates its corresponding part l_k = $l_{k1}l_{k2}...l_{kp_k}$ in the MWE lemma l. It is based on a character-based conditional GRU augmented with an attention mechanism (Bahdanau et al., 2014). Every w_k is encoded as a vector which is the concatenation of the following features: its context-free character-based embedding $C(w_k)$, its left context[5] h_{k-1} in the MWE (h_{k-1} being the output of the GRU at time stamp $k-1$), a tag TAG_k, its position k and the MWE length n. TAG_k is the embedding of the predicted POS tag of w_k, sometimes concatenated with the embedding of the gold MWE POS tag.

Our model has some limitations. First, the input form and the produced base form must have the same number of words. Secondly, the sequential nature of the model and the one-to-one correspondance are not very adequate to model lemmatization modifying the word order. For instance, the lemmatization of the verbal expression *decision [was] made* in the passive form involves word

[3] Preliminary experiments showed that bidirectional GRUs display lower results than left-to-right GRUs, which might be explained by the fact that we deal with suffix-based inflectional languages.

[4] Preliminary experiments showed that using bidirectional GRUs has no positive effect.

[5] The use of contextualized embedding h_k of the current word instead of h_{k-1} has shown slightly lower results.

Lang	Type	Source	Set	Nb of MWEs	Nb of ≠ MWEs	Nb of MWE POS	Nb of simple words	Nb ofs unk. MWE
FR	Dict	DELA (Silberztein, 1994)/Morphalou (ATILF, 2016)	Train	118346	104938	11	956834	
			Dev	4627	4335	10		4342 (93.7%)
			Test	4163	3956	11		3975 (95.3%)
	Corpus	FTB (Abeillé et al., 2003; Seddah et al., 2013)	Train	12373	2948	9	456833	
			Dev	1227	667	9		142 (11.6%)
			Test	1835	890	9		234 (12.8%)
	Corpus	PARSEME Shared Task 1.0 (ST) (Ramisch et al., 2018; Candito et al., 2017; Pasquer et al., 2018)	Train	3461	1901	1	0	
			Dev	486	327	1		451 (92.8%)
			Test	491	328	1		333 (67.8%)
PL	Dict	SEJF (Gralinski et al., 2010) SEJFEK (Savary et al., 2012)	Train	206471	121816	6	0	
			Dev	4252	2800	4		4250 (100.0%)
			Test	4909	3181	3		4819 (98.2%)
	Corpus	KPWr 1.2 (Broda et al., 2012)	Train	2862	1864	1	33274	
			Dev	912	805	1		273 (29.9%)
			Test	987	824	1		303 (30.7%)
IT	Dict	Unitex dictionary (Vietri and Elia, 2000)	Train	30415	29620	1	0	
			Dev	993	959	1		992 (99.9%)
			Test	997	979	1		997 (100%)
PT	Dict	Unitex dictionary (Ranchhod et al., 1999)	Train	8996	8681	2	0	
			Dev	997	964	2		994 (99.7%)
			Test	997	958	2		995 (99.8%)
BR	Dict	Unitex dictionary (Muniz et al., 2005)	Train	2987	2959	3	0	
			Dev	483	476	2		483 (100%)
			Test	497	492	2		497 (100‘%)

Table 1: Dataset sources and statistics. The column *Nb of ≠ MWEs* refers to the number of MWE types (i.e. number of different MWEs). The column *Nb of MWE POS* refers to the size of the set of MWE-level POS tags

reordering, namely *make decision*.

3 Dataset

Our dataset[6] embodies sets of gold pairs (MWE form, MWE lemma) in five languages namely Brazilian Portuguese (BR), French (FR), Italian (IT), Polish (PL), Portuguese Portuguese (PT). It includes both token-based and type-based data. Token-based data are derived from annotated corpora and are intended to be used to evaluate our approach on a real MWE distribution. Type-based data are derived from different morphosyntactic dictionaries and are intended to be used to evaluate the coverage and robustness of our approach. They are divided in train/dev/test splits. Table 1 displays the dataset sources and statistics. French and Polish data are by far the larger datasets and includes both token- and type-based resources. Italian and Portuguese data are smaller and only type-based. They are derived from the freely available dictionaries in the Unitex plateform (Paumier et al., 2009). We constructed our dataset by applying some automatic preprocessing to resolve tokenization and lemma discrepancies between the different sources, and to filter MWEs whose number of words is not equal to the number of words of the lemma, since our approach is based on a word-to-word process (1.6% of the MWEs are thus taken off in French). For token-based datasets, we used the official splits used in Ramisch et al. (2018) and Seddah et al. (2013) for French, and in Marcińczuk (2017) for Polish. For dictionary-based resources, we applied a random split by taking care of keeping all entries with the same lemma in the same split.

For every language, we constructed a unique[7] training set composed of the different train parts of the different resources used. We also augmented our training sets with gold pairs (simple-word form, simple-word lemma) to account for simple-word lemmatization knowledge in the MWE lemmatization process. This information comes from the same sources as MWEs.

	Dev (MWEs)		Test (MWEs)		Test (words)	
	all	unk.	all	unk.	all	unk.
FR ftb	95.9	91.5	95.6	93.2	98.0	96.8
FR shared task	73.1	73.1	75.2	75.2	82.7	82.6
FR dict	86.0	86.9	87.5	88.4	89.9	91.1
PL corpus	88.9	75.5	88.9	75.5	94.1	87.7
PL dict	59.5	59.5	58.6	59.0	76.8	76.8
IT	91.7	91.7	91.7	91.7	92.9	92.9
PT	89.7	89.7	88.2	88.4	95.1	95.1
BR	84.6	84.6	81.6	81.6	90.6	90.6

Table 2: Final results for all and unknown MWEs. Columns *Dev(MWEs)* and *Test(MWEs)* provide MWE-based accuracy on the dev and test sets respectively. Column *Test(words)* gives word-based accuracy on the test set.

[6] Datasets and code can be found at the following url: `https://git.atilf.fr/parseme-fr/deep-lexical-analysis`. Note that the French Treebank data are distributed upon request because of license specificities.

[7] For French, ST data train set was separated from the rest.

4 Experiments

Experimental setup. We manually tuned the hyperparameters of our system on the dev sections. Our final results on test sections were obtained using the best hyperparameter setting for the dev sections (hidden layer size: 192, character embedding size: 32, tag embedding size: 8, learning rate: 0.005, dropout: 0.25). We used UDPipe (Straka and Straková, 2017) to predict word POS tags for all languages. We also included predicted morphological features for Polish.

Evaluation metrics. We evaluated our system by using two metrics: MWE-based accuracy and word-based accuracy. MWE-based accuracy, also used for tuning, accounts for the proportion of MWEs that have been correctly lemmatized. Word-based accuracy indicates the total proportion of words that have been given the correct corresponding lemma part.

Results. Table 2 displays our final results on the dev and test sets of our five languages. First, it shows that our system generalizes well on unknown MWEs (columns *unk.*). For type-based data, scores on unknown MWEs are comparable or slightly better than for all MWEs. For token-based data, the MWE-based accuracy loss is reasonable, ranging from almost 0 point for French verbal expressions (ST data) to 13 points for Polish MWEs. Our system shows good performances on French. On similar languages (BR, IT, PT), results are lower, but rather good given the limited size of the training sets. The system shows disappointing results for Polish, especially for the dictionary. On the token-based dataset, results are very far from the ones obtained by the rule-based system of (Marcińczuk, 2017) which displays around 98% accuracy using 27 rules and dictionary information. Polish being a morphologically-rich language, the encoding of morphological constraints would deserve more investigations. The system also shows lower scores for verbal expressions in French, which show much morphological and syntactic variation.

We also evaluated our system to lemmatize simple words, as it would have been convenient to have a single system processing the lemmatization on both simple words and MWEs. However, it did not show satisfying results: we obtained a score of 73% on the FTB corpus, against 99% when the system is trained on simple words only.

	Dict	FTB
Complete system	86.0	95.9
- GRU on word sequence	75.6	88.1
- word POS tags	81.9	95.7
- position and length feats	83.6	95.8
- simple words in train set	78.3	88.9
Complete system + MWE gold tag	90.0	97.1
baseline UDPipe adaptation	83.5	95.5
baseline word-to-word	54.0	73.0

Table 3: MWE-based accuracy on dev section for French with different architectures and comparison with baselines.

5 Discussion

Ablation study. In order to evaluate the impact of the different components of our neural architecture, we performed an ablation study on French, by removing (1) the GRU component on the word sequence, (2) the word POS tags, (3) word position and MWE length information, (4) simple-word examples from train set. Table 3 displays the results on the dev section of the French data excluding the ST data. The GRU component appears crucial to capture morphosyntactic constraints (8-10 point gain). The use of simple-word lemmatization knowledge has also a significant impact (7-8 point gain). Word POS tags are mainly beneficial for the dictionary evaluation (4-point gain). We also evaluated the impact of adding the gold MWE POS, which are mainly beneficial in a dictionary evaluation setting (4-point gain).

	Our system	Baseline
FR ftb	95.9	95.5
FR dict	86.0	83.5
PL corpus	88.9	70.1
PL dict	59.5	46.5

Table 4: Best result for our system compared to UDPipe adaptation baseline for French and Polish dev sets. The table shows MWE-based accuracy.

Comparison with baselines. We compared our system with two baselines, both using UDPipe (Straka and Straková, 2017).
The first one consists in training UDPipe in a special way. More precisely, it is trained on sequences of simple words of the train corpora, plus on the MWE word sequences of the training data set. In order to give cues about the MWE internal structure to UDPipe, we provide MWE words with IOB-like tags indicating their relative positions in the MWE, in addition to their POS-tags/MWE-tag, in the train set. For instance, the French MWE

cartes bleues (lit. cards blue, tr. credit cards) would be annotated in the following way (with POS-tags): *cartes/carte/B-NOUN bleues/bleue/I-ADJ*.

The second one simply consists in lemmatizing each word of the MWE separately, with UDPipe already trained with the basic UD model. The output MWE lemma is the concatenation of the predicted lemmas of all MWE words. Table 3 shows that this baseline is not competitive with respect to the UDPipe adaptation baseline.

Table 4 compares the performances of our system with the best baseline on dev datasets[8] for French and Polish. The baseline consistently shows lower scores for Polish and French. The best baseline ranges from 0.4-to-2.5-point loss for French and more than 10-point loss for Polish.

	French		Polish	
	Dict	Corp	Dict	Corp
(a) MWE lemma = MWE form	94.2 *(65.0)*	97.9 *(83.2)*	74.5 *(12.7)*	93.3 *(54.8)*
(b) MWE lemma = concat(lemmas)	95.8* *(55.8)*	99.4 *(70.4)*	67.4* *(28.5)*	90.9 *(43.1)*
Union of (a) and (b)	93.1 *(84.1)*	97.8 *(95.2)*	68.1 *(38.2)*	91.6 *(66.0)*
Intersection of (a) and (b)	99.1 *(35.2)*	100.0 *(62.5)*	85.5 *(3.0)*	93.4 *(31.9)*
Other MWE	82.5 *(15.9)*	85.7 *(4.8)*	57.3 *(61.8)*	83.2 *(34.0)*

Table 5: MWE-based accuracy on dev section according to MWE subclasses. * indicates that lemmas were predicted by UDPipe. Otherwise they are gold. Numbers between parentheses indicate the repartition of the MWE subclasses in the tested dataset (in percentage).

Results by MWE subclasses. Table 5 compares results for different lemmatization cases for French and Polish on dev data: the MWE lemma corresponds to (1) the MWE form, (2) the concatenation of the word lemmas, (3) other cases. In French, our system performs rather well on the second case. In Polish, system performs better on the first case. It is worth noticing that our system performs very well on MWEs that belong to both cases (1) and (2), especially for French. There is a significant gap in performances with the other cases for both languages. Note that the proportion of MWEs belonging to the other cases is much greater in Polish than in French. This might partially explains why the system performs so poorly on Polish data.

6 Related work

Lemmatization of simple words has already received much attention. Recently, researchers proposed approaches based on statistical classification, like predicting edit tree operations transforming word forms into lemmata (Grzegorz Chrupala and van Genabith, 2008; Müller et al., 2015) or predicting lemmatization rules consisting in removing and then adding suffixes and prefixes (Straka and Straková, 2017). Using the deep learning paradigm, Schnober et al. (2016) and Bergmanis and Goldwater (2018) proposed attention-based encoder-decoder lemmatization.

Regarding multiword lemmatization, Oflazer and Kuruoz (1994) and Oflazer et al. (2004) historically proposed to perform finite-state rule-based morphology analysis. More recently, the task was mainly investigated for highly inflectional languages like Slavic ones. Research focused mainly on approaches based on heuristics (Stankovic et al., 2016; Marcińczuk, 2017), string distance metrics (Piskorski et al., 2007) and tagging (Radziszewski, 2013).

7 Conclusion

In this paper, we presented a novel architecture for MWE lemmatization relying on a word-to-word process based on a deep encoder-decoder neural network. It uses both the morphological information of the individual words and their internal context in the MWE. Evaluations for five languages showed that the proposed system generalizes well on unknown MWEs, though results are disappointing for a language with very rich morphology like Polish and for verbal expressions. This would require further more detailed investigation. Another line of research for future work would consist in integrating transformers in our system and in evaluating it on more languages.

Acknowledgment

This work has been partly funded by the French *Agence Nationale pour la Recherche*, through the PARSEME-FR project (ANR-14-CERA-0001). It was also supported by the French PIA project *Lorraine Université d'Excellence*, reference ANR-15-IDEX-04-LUE, via the Lex4Know project. Authors are grateful to the anonymous reviewers for their insightful comments. Authors would also like to thank Caroline Pasquer for her dataset of normalized French verbal expressions.

[8] Results on test sets show the same trend.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks*. Kluwer, Dordrecht.

ATILF. 2016. Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Bartosz Broda, Michal Marcinczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardynski. 2012. Kpwr: Towards a free corpus of polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3218–3222.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer, and Jean-Yves Antoine. 2017. Annotation d'expressions polylexicales verbales en français. In *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Actes de TALN, volume 2 : articles courts, pages 1–9, Orléans, France.

Mathieu Constant, Glen Eryiit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Filip Gralinski, Agata Savary, Monika Czerepowicka, and Filip Makowiecki. 2010. Computational lexicography of multi-word units. how efficient can it be? In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 2–10, Beijing, China. Coling 2010 Organizing Committee.

Maurice Gross. 1986. Lexicon grammar. the representation of compound words. In *Proceedings of the 11th International Conference on Computational Linguistics, COLING '86, Bonn, Germany, August 25-29, 1986*, pages 1–6.

Georgiana Dinu Grzegorz Chrupala and Josef van Genabith. 2008. Learning morphology with morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Michał Marcińczuk. 2017. Lemmatization of multiword common noun phrases and named entities in polish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 483–491. INCOMA Ltd.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Marcelo C.M. Muniz, Maria V. Nunes das Graas, and Eric Laporte. 2005. Unitex-pb, a set of flexible language resources for brazilian portuguese. In *Proceedings of the Workshop on Technology on Information and Human Language (TIL)*, pages 2059–2068.

Kemal Oflazer, Özlem Çetinoğlu, and Bilge Say. 2004. Integrating morphology with multi-word expression processing in turkish. In *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 64–71, Barcelona, Spain. Association for Computational Linguistics.

Kemal Oflazer and Ilker Kuruoz. 1994. Tagging and morphological disambiguation of turkish text. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 144–149, Stuttgart, Germany. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2018. If you've seen some, you've seen them all: Identifying variants of multiword expressions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2582–2594. Association for Computational Linguistics.

Paumier, Nakamura, and Voyatzi. 2009. Unitex, a corpus processing system with multi-lingual linguistic resources. In *eLexicography in the 21st century: new challenges, new applications (eLEX'09)*, pages 173–175.

Jakub Piskorski, Marcin Sydow, and Anna Kup. 2007. Lemmatization of Polish Person Names. In *ACL 2007. Proceedings of the Workshop on Balto-Slavonic NLP 2007*, pages 27–34. Association for Computational Linguistics.

Adam Radziszewski. 2013. Learning to lemmatise polish noun phrases. In *Proceedings of the 51st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 701–709, Sofia, Bulgaria. Association for Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Gngr, Abdelati Hawwari, Uxoa Iurrieta, Jolanta Kovalevskait, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartn, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Elisabete Ranchhod, Cristina Mota, and Jorge Baptista. 1999. A computational lexicon of portuguese for automatic text parsing. In *Proceedings of SIGLEX'99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL*, pages 74–81.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

Agata Savary, Bartosz Zaborowski, Aleksandra Krawczyk-Wieczorek, and Filip Makowiecki. 2012. Sejfek - a lexicon and a shallow grammar of polish economic multi-word units. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 195–214, Mumbai, India. The COLING 2012 Organizing Committee.

Carsten Schnober, Steffen Eger, Erik-Lân Do Dinh, and Iryna Gurevych. 2016. Still not there? comparing traditional sequence-to-sequence models to encoder-decoder neural networks on monotone string translation tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1703–1714, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Max D. Silberztein. 1994. Intex: A corpus processing system. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994), pages 579–583, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ranka Stankovic, Cvetana Krstev, Ivan Obradovic, Biljana Lazic, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

S. Vietri and A. Elia. 2000. Electronic dictionaries and linguistic analysis of italian large corpora. In *JADT 2000 - Actes des 5es Journees internationales d'Analyse statistique des Donnes Textuelles*.