A Framework for Understanding the Role of Morphology in Universal Dependency Parsing

Mathieu Dehouck, Pascal Denis

A Framework for Understanding the Role of Morphology in Universal Dependency Parsing

Mathieu Dehouck
Univ. Lille, CNRS, UMR 9189 - CRIStAL
Magnet Team, Inria Lille
59650 Villeneuve d'Ascq, France
mathieu.dehouck@inria.fr

Pascal Denis
Magnet Team, Inria Lille
59650 Villeneuve d'Ascq, France
pascal.denis@inria.fr

Abstract

This paper presents a simple framework for characterizing morphological complexity and how it encodes syntactic information. In particular, we propose a new measure of morpho-syntactic complexity in terms of governor-dependent preferential attachment that explains parsing performance. Through experiments on dependency parsing with data from Universal Dependencies (UD), we show that representations derived from morphological attributes deliver important parsing performance improvements over standard word form embeddings when trained on the same datasets. We also show that the new morpho-syntactic complexity measure is predictive of the gains provided by using morphological attributes over plain forms on parsing scores, making it a tool to distinguish languages using morphology as a syntactic marker from others.

1 Introduction

While word embedding has proven a good solution to reduce data sparsity in parsing (Koo et al., 2008), treating word forms as atomic units is at odds with the fact that words have a potentially complex internal structure. Furthermore, it makes parameters estimation difficult for morphologically rich languages (MRL) in which the number of possible forms a word can take can be very large[1].

Recently, researchers have started to work on morphologically informed word embeddings (Cao and Rei, 2016; Botha and Blunsom, 2014), aiming at better capturing both lexical, syntactic and morphological information. But encoding lexicon and morphology in the same space makes it difficult to distinguish the role of each in syntactic tasks such as dependency parsing. Furthermore, morphologically rich languages for which we hope to see a real impact from those morphologically aware representations, might not all rely to the same extent on morphology for syntax encoding. Some might benefit mostly from reducing data sparsity while others, for which paradigm richness correlate with freer word order (Comrie, 1981), will also benefit from morphological information encoding.

This paper aims at characterizing the role of morphology as a syntax encoding device for various languages. Using simple word representations, we measure the impact of morphological information on dependency parsing and relate it to two measures of language morphological complexity: the basic form per lemma ratio and a new measure (*HPE*) defined in terms of head attachment preference encoded by its morphological attributes. We show that this new measure is predictive of parsing result differences observed when using different word representations and that it allows one to distinguish amongst morphologically rich languages, those that use morphology for syntactic purpose from those using morphology as a more semantic marker. To the best of our knowledge, this work is the first attempt at systematically measuring the syntactic content of morphology in a multi-lingual environment.

Section 2 presents the representation learning method and the dependency parsing model. It also defines two measures of morphological complexity. Section 3 describes the experimental setting and analyses parsing results in terms of the previously defined morphological complexity measures. Section 4 gives some conclusions and future work perspectives.

[1] A typical English noun has 2 forms while a Finnish one may have more than 30. This shows in data as English lemmas have 1.39 forms on average while Finnish ones have 2.19, as measured on UD data (Nivre et al., 2016).

2 Framework

This section details: (i) our method for learning lexical and morphological representations, (ii) how these can be used for graph-based dependency parsing, and (iii) how to measure morphological complexity. Our representation learning and parsing techniques are purposely very simple in order to let us separate lexical and morphological information and weight the role of morphology in dependency parsing of MRL.

2.1 Word Representation

We construct separate vectorial representations for lemmas, forms and morphological attributes, either learned via dimension reduction of their own cooccurrence count matrices or represented as raw one-hot vectors.

Let $\mathcal{V}$ be a vocabulary (it can be lemmas or forms or morphological attributes (incl. values for POS, number, case, tense, mood...)) for a given language. Correspondingly, let $\mathcal{C}$ be the set of contexts defined over elements of $\mathcal{V}$. That is, lemmas appear in the context of other lemmas, forms in the context of forms, and attributes in the context of attributes. Then, given a corpus annotated with lemmas and morphological information, we can gather the cooccurrence counts in the matrix $\boldsymbol{M} \in \mathbb{N}^{|\mathcal{V}|\times|\mathcal{C}|}$, such that $\boldsymbol{M}_{ij}$ is the frequency of lemma (form or morphological attributes) $\mathcal{V}_i$ appearing in context $\mathcal{C}_j$ in the corpus. Here, we consider plain sequential contexts (i.e. surrounding bag of "words") of length 1, although we could extend them to more structured contexts (Bansal et al., 2014). Those cooccurrence matrices are then reweighted by unshifted Positive Point-wise Mutual Information (PPMI) and reduced via Singular Value Decomposition (SVD). For more information on word embedding via matrix factorization, please refer to (Levy et al., 2015).

Despite its apparent simplicity, this model is as expressive as more popular state of the art embedding techniques. Indeed, Goldberg and Levy (2014) have shown that the SkipGram objective with negative sampling of Mikolov's Word2vec (2013) can be framed as the factorization of a shifted PMI weighted cooccurrence matrix.

This matrix reduction procedure gives us vectors for lemmas, forms and morphological attributes, noted $\boldsymbol{R}$. Note that while a word has only one lemma and one form, it will often realize several morphological attributes. We tackle this issue by simply summing over all the attributes of a word (noted $Morph(w)$). If we note $\boldsymbol{r}_w$ the vectorial representation of word w we have:

$$\boldsymbol{r}_w = \sum_{a \in Morph(w)} \boldsymbol{R}_a.$$

Simple additive models have been shown to be very efficient for compositionally derived embeddings (Arora et al., 2017).

2.2 Dependency Parsing

We work with graph-based dependency parsing, which offers very competitive parsing models as recently re-emphasized by Dozat et al. (2017) in the CONLL 2017 shared-task on dependency parsing (Zeman et al., 2017).

Let $x = (w_1, w_2, ..., w_n)$ be a sentence, $\mathcal{T}_x$ be the set of all possible trees over it, $\hat{y}$ the tree that we predict for x, and $Score(\bullet, \bullet)$ a scoring function over sentence-tree pairs :

$$\hat{y} = \underset{t \in \mathcal{T}_x}{\text{argmax}}\ Score(x, t).$$

We use edge factorization to make the inference problem tractable. A tree score is thus the sum of its edges scores. We use a simple linear model:

$$Score(x, t) = \sum_{e \in t} \boldsymbol{\theta}^\top \cdot \boldsymbol{\phi}(x, e),$$

where $\boldsymbol{\phi}(x, e)$ is a feature vector representing edge e in sentence x, and $\boldsymbol{\theta} \in \mathbb{R}^m$ is a parameter vector to be learned.

The vector representation of an edge e_{ij} whose governor is the i-th word w_i and dependent is the j-th word w_j, is defined by the outer product of their respective representations in context. Let $\oplus$ note vector concatenation, $\otimes$ the outer product and $w_{k\pm1}$ be the word just before/after w_k, then: $\boldsymbol{v}_i = \boldsymbol{w}_{i-1} \oplus \boldsymbol{w}_i \oplus \boldsymbol{w}_{i+1}$, $\boldsymbol{v}_j = \boldsymbol{w}_{j-1} \oplus \boldsymbol{w}_j \oplus \boldsymbol{w}_{j+1}$ and

$$\boldsymbol{\phi}(x, e_{ij}) = \text{vec}(\boldsymbol{v}_i \otimes \boldsymbol{v}_j)\ \in \mathbb{R}^{9d^2}.$$

Recall that $\boldsymbol{w}_i$ of length $d \ll \mathcal{V}$ is a vector from $\boldsymbol{R}$.

We use the averaged Passive-Aggressive online algorithm for structured prediction (Crammer et al., 2006) for learning the model $\boldsymbol{\theta}$. Given a score for each edge, we use Eisner algorithm (Eisner, 1996) to retrieve the best projective spanning tree. Even though some languages display a fair amount of non-projective edges, on average Eisner algorithm scores higher than Chu-Liu-Edmonds algorithm (Chu and Liu, 1965) in our setting.

2.3 Measuring Morpho-Syntactic Complexity

Some languages use morphological cues to encode syntactic information while other encode more semantic information with them. For example, the Case feature (especially core cases) is of prime syntactic importance, for it encodes the type of relation words have with each other. On the contrary, the Possessor feature (in Hungarian for example) is more semantic in nature and need not impact sentence structure. This remark would support different treatment for each language. However, those languages tend to be treated equally in works dealing with MRL.

Form to Lemma Ratio A basic measure of morphological complexity is the form per lemma ratio, we note it F/L. It captures the tendency of words to inflect in a given language. Because some word classes tend not to inflect and not all forms are equally productive, we note F/iL the ratio of form per inflected lemma. Given a language l with a lemma vocabulary $\mathcal{V}^l$ and a form counting function $c : \mathcal{V}^l \rightarrow \mathbb{N}$ that returns the number of forms a lemma can take, we have:

$$F/L(l) = \frac{1}{|\mathcal{V}^l|} \sum_{w \in \mathcal{V}^l} c(w),$$

$$F/iL(l) = \frac{1}{|\mathcal{V}_i^l|} \sum_{w \in \mathcal{V}_i^l} c(w), \mathcal{V}_i^l = \{w \in \mathcal{V}^l | c(w) > 1\}$$

F/L and F/iL do not measure the informative content of morphology, but simply its productivity. Bentz et al. (2016) compared five different measures of morphological complexity amongst which word entropy and the micro-averaged version of F/L (they call it TTR) and showed that they all have high positive correlation given enough data.

Head POS Entropy In order to compare the morpho-syntactic complexity of different languages, we introduce a new measure called *Head Part-of-speech Entropy* or *HPE*. The *HPE* of a token t represents the amount of information t has about the part-of-speech of its governor. More formally, let $POS(Gov(t))$ be the set of parts-of-speech that t can depend on, and let $\pi_t(p)$ be the probability of t actually depending on part-of-speech p, then the *HPE* is defined as:

$$HPE(t) = \sum_{p \in POS(Gov(t))} -\pi_t(p) log_2(\pi_t(p)).$$

This is a measure of a token preferencial attachment to its head. A token with a low *HPE* tends to attach often to the same part-of-speech, while a token with a high *HPE* will attach to many different parts-of-speech. Thus a language with a low *HPE* will tend to encode a lot of syntactic information in the morphology, rather than in word order say.

For example, a noun can attach to another noun like a genitive, or to a verb as a subject or object, or even to an adjective in the case of transitive adjective. French nouns do not inflect for case, thus attachment to another noun or verb can only be infered from words relative positions. On the contrary, Gothic nouns do inflect for case, thus making verb or noun attachment clear directly from the morphological analysis.

We compute the *HPE* of a language as the averaged *HPE* of its attributes sets over a given corpus. Likewise, we use the empirical counts as a surrogate for c in F/L and F/iL.

3 Experiments

In order to test the hypothesis that morphological representations contain syntactic information crucial for dependency parsing of morphologically rich languages, but that this information is not equally distributed across MRL, we run experiments on data from the Universal Dependencies (Nivre et al., 2016) project.

Data Description For conciseness, we focused on eleven languages that display varying degrees of morphological complexity and belong to four different language families. Basque (eu) is an isolate and it is an ergative language. English (en), Gothic (got), Danish (da) and Swedish (sv) are Germanic languages, and French (fr) and Romanian (ro) are Romance languages (Indo-European). Finnish (fi), Estonian (et) and Hungarian (hu) are Finno-Ugric languages. Hebrew (he) is a Semitic language. Basic statistics are provided in Table 1.

Experimental Settings For the experiments we use the train/dev/test data provided by UD 2.0. Basic statistics about the data are reported in the appendix. Lemmas and forms are embedded in 150 dimensions, while Morphological attributes are embedded in 50 dimensions, because they are much less numerous (less than 100). All embeddings are induced on their language respective train set only using a context window of size 1 (i.e. the

	da	en	et	eu	fi	fr	got	he	hu	ro	sv
Train	4383	12543	2263	5396	12217	14553	3387	5241	910	8043	4303
POS	17	17	16	16	15	17	14	16	16	17	16
Feats	44	35	58	69	88	36	40	48	73	59	39
F/L	1.44	1.39	1.60	2.32	2.19	1.38	2.44	1.83	1.46	2.03	1.59
F/iL	2.80	2.76	3.35	4.29	4.68	3.15	4.20	3.39	3.03	3.76	2.91
HPE	1.07	1.12	0.55	0.51	0.57	0.87	0.60	1.01	0.60	0.71	0.84

Table 1: Basic datasets statistics. The first line gives the number of train sentences for each language. The second and third give the number of part-of-speech and of morphological attribute values for each language. The fourth and fifth lines reports forms per lemma ratios. Last line gives the HPE.

		da	en	et	eu	fi	fr	got	he	hu	ro	sv
Lem	OH	58.47	67.05	44.96	60.27	56.06	70.93	60.63	65.70	47.32	68.69	61.30
	Emb	68.33	75.21	59.23	69.59	66.90	71.90	71.14	72.52	51.04	72.83	73.27
Form	OH	56.92	65.83	41.36	57.67	51.50	70.35	58.68	67.08	44.35	67.2	58.54
	Emb	70.64	75.13	57.36	65.64	60.38	76.05	68.72	72.68	55.06	73.21	72.72
Morph	OH	**73.76**	**76.27**	**71.21**	**73.81**	**75.58**	78.67	**76.88**	**77.65**	**69.67**	**76.57**	**76.42**
	Emb	73.27	76.17	70.20	73.21	73.33	**78.79**	76.37	76.95	69.38	76.32	76.02

Table 2: UAS scores for parsers using lemmas (Lem), word forms (Form) or morphological attributes (Morph) representations as features. For each type, we report results using one-hot representation (OH) and results using embeddings (Emb).

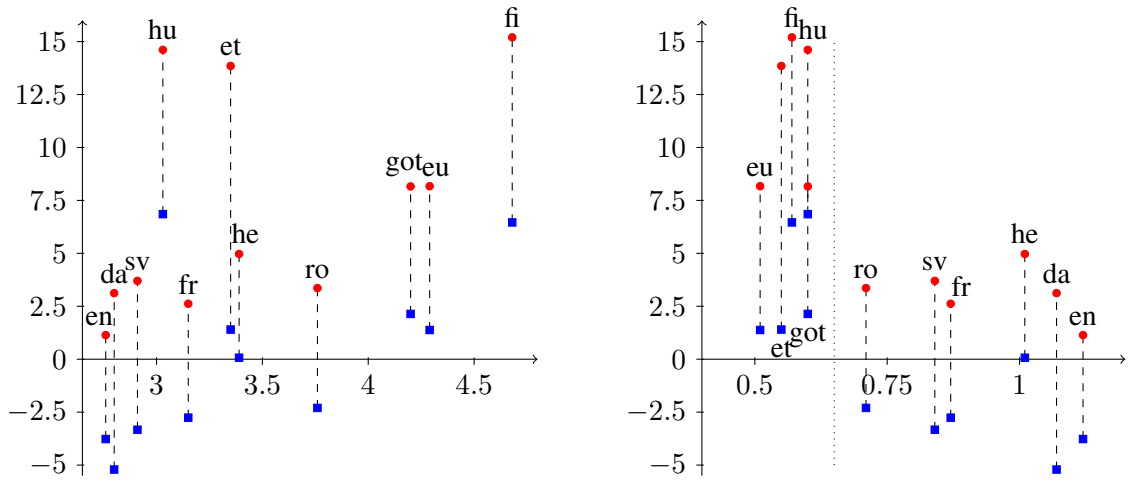
directly preceding and following words).

Parsers are trained for 10 iterations using either lemma, form or morphological representations, and we pick the best iteration on the basis of UAS on the development set.

While we used gold lemmas as provided in the corpora, we ran two experiments for morphological attributes, one with gold attributes and one with predicted attributes. Morphological attributes are predicted with a simple multinomial logistic regression per attribute (POS, Tense, Case, Gender...), where we add a special *undef* value (except for POS) to represent the lack of an attribute (e.g., nouns have no Tense in English). The models predict attribute values for the center word of trigrams represented by feature vectors encoding word prefixes and suffixes of length 1, 2 and 3, word length and capitalization. We used the logistic regression implemented in the Scikit-Learn (Pedregosa et al., 2011) library with the default settings. It can output an argmaxed decision or a softmaxed decision, thus we tried both as input to the parser. The argmaxed decision gives a vector of zeros and ones, while the softmaxed decision gives a continuous vector with each each attributes summing to one (the probability assigned to each possible value for Gender like Masculine, Feminine, Neuter and Undef must sum to one). Then those vectors are used unchanged for the one-hot representation or passed through an embedding matrix for the embedding representation.

Results For clarity, we focus on comparing results using form embeddings and gold morphological representations. They are given in Table 2. Because the analysis carries to the labeled case, we stick to unlabeled scores (UAS) for the analysis. A more complete table is provided in the appendix as well as a complete labeled accuracy score (LAS) table. Morphological complexity measures are also reported.

One-hot gold morphological attributes consistently outerperform form embeddings. This is expected since forms embedding were trained on much fewer data than usually considered necessary. However, improvements are not consistent across languages, ranging from 1.14 point for English to 15.20 points for Finnish. While those differences are not explained by morphological productivity alone (Figure 1a), a measure of preferential attachment gives a good account of them (Figure 1b). Those inconsistencies become even more striking, considering results using predicted attributes. We notice that despite a general drop of performance of 5-12 points, predicted attributes



(a) F/iL

(b) HPE

Figure 1: Accuracy differences (y-axis) between parsers using form embeddings and parsers using one-hot attributes, with respect to morphological complexity (x-axis). Red dots represent the gold attributes scores and blue squares the predicted attributes scores.

still perform significantly better than form embeddings for those morphologically rich languages that have an *HPE* lower than 0.65 as depicted on Figure 1b.

Figures 1a and 1b plot the differences in parsing scores. For each language, the red dot corresponds to the score difference between using form embeddings and gold attributes one-hot representations, and the blue square corresponds to the score difference between using the same form embeddings and predicted attributes softmax representations (the complete scores are given in the appendix). Figure 1a plots those differences with regard to the form per inflected lemma ratio (F/iL) and Figure 1b plots those differences with regard to the head POS entropy (*HPE*).

Both Figures show trends. Score differences seem to increase with F/iL and decrease with *HPE*. But while the F/iL plot suffers outliers (Hungarian, Estonian and Romanian), the *HPE* plot shows a clear boundary between languages benefiting fully from morphological information (even predicted) and those benefiting primarily from reducing data sparsity. While Hebrew seems to be an outlier, it might be due to its annotation style, where attached prepositions, articles and possessive markers are treated as independent words rather than morphological inflection as other languages do, thus artificially increasing the parsing accuracy with a lot of trivial dependencies.

This shows that indeed, *HPE* is a good measure of the syntactic informativeness of a language morphology, and that it can help deciding between encoding morphological information or just reducing data sparsity. Furthermore, it seems to be link to the distinction that Kibort and Corbett (2010) do between morphosyntax and morphosemantic.

4 Conclusion

We have contributed a new measure of morpho-syntactic complexity (*HPE*) that helps distinguishing languages that use morphology for syntactic purpose from languages that use morphology to encode more semantic information. We showed that this measure correlates much more with differences in parsing results using morphological representations than the simple form per lemma ratio. It could thus be used to help designing language specific word representations.

It is worth mentioning that we focused here on dependent marked head selection. It would be interesting to have a similar measure for head-marking situations with dependencies marked on the governor. We leave it for future work.

5 Acknowledgement

This work was supported by ANR Grant GRASP No. ANR-16-CE33-0011-01 and Grant from CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020. We also thank the reviewers for their valuable feedback.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *International Conference on Learning Representations 2017*.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815. Association for Computational Linguistics.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic. 2016. A comparison between morphological complexity measures: typological data vs. language corpora. In *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pages 142–153.

Jan A. Botha and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. *CoRR*, abs/1405.4273.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. *CoRR*, abs/1606.02601.

Y. J. Chu and T. H. Liu. 1965. On the shortest arborescence of a directed graph. *Science Sinica*, 14.

Bernard Comrie. 1981. *Language universals and linguistic typology : syntax and morphology / Bernard Comrie*. Blackwell, Oxford :.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Jason Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

A Kibort and GG Corbett. 2010. *Features: perspective on a key notion in linguistics.* Oxford University Press, Oxford.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3:211–225.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 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. Association for Computational Linguistics.

		da	en	et	eu	fi	fr	got	he	hu	ro	sv
	POS	87.37	87.24	84.49	86.09	86.14	90.60	90.44	90.93	88.43	90.64	89.23
	Attributes	98.02	97.74	96.20	97.56	97.09	97.60	96.00	97.84	92.76	97.76	96.88
	Form Emb	**70.64**	**75.13**	57.36	65.64	60.38	**76.05**	68.72	72.68	55.06	**73.21**	**72.72**
Morph	Hard OH	64.69	69.32	57.16	64.51	64.33	72.82	69.94	71.60	61.80	71.36	67.99
Morph	Soft OH	65.43	71.36	**58.76**	**67.02**	**66.84**	73.29	**70.86**	**72.75**	**61.91**	72.03	69.48
Morph	Hard Emb	64.19	69.51	55.53	64.10	62.58	72.28	69.29	71.51	59.62	70.86	67.82
Morph	Soft Emb	65.33	70.75	57.24	66.18	65.04	73.18	70.46	71.85	60.64	71.34	68.60

Table 3: UAS scores for parsers using predicted morphological attributes. The two first rows are POS and averaged attributes prediction accuracy. The third row reports UAS using form representations for comparison purpose. Rows 4 to 7 give UAS using morphological representations, either one-hot or embedding. Regressors output a probability distribution per morphological feature, we either use those soft decision as input for the parser (Soft) or apply argmax first (Hard).

		da	en	et	eu	fi	fr	got	he	hu	ro	sv
Lem	OH	48.09	57.09	25.30	45.96	40.78	64.88	46.85	54.91	27.80	56.89	48.61
Lem	Emb	62.47	70.95	48.17	62.52	59.34	65.62	61.37	64.41	41.59	64.76	65.70
Form	OH	45.12	54.97	21.29	40.53	34.59	61.95	45.19	55.82	25.60	53.83	45.00
Form	Emb	65.09	71.20	45.79	57.42	52.67	70.81	59.35	66.92	44.30	65.13	64.93
Morph	OH	**69.19**	**72.32**	**64.06**	**68.19**	**71.00**	73.92	**71.04**	**72.66**	**64.31**	**68.94**	**69.97**
Morph	Emb	68.71	72.22	62.81	67.30	68.70	**73.96**	70.41	71.77	63.45	68.76	69.69

Table 4: LAS scores for parsers using lemmas (Lem), forms (Form) or morphosyntactic attributes (Morph) representations as features. Representations are either embeddings or one-hot.

		da	en	et	eu	fi	fr	got	he	hu	ro	sv
	Form Emb	**65.09**	**71.20**	45.79	57.42	52.67	**70.81**	59.35	**66.92**	44.30	**64.13**	**65.93**
Morph	hard OH	58.33	62.64	43.80	55.81	54.42	66.66	59.73	63.74	52.41	62.10	60.29
Morph	soft OH	59.68	65.59	**47.05**	**59.43**	**58.74**	67.39	**62.36**	66.25	**53.63**	63.26	62.44
Morph	hard Emb	57.72	62.73	42.22	55.06	52.79	66.25	59.14	63.57	49.99	61.67	60.03
Morph	soft Emb	59.13	64.97	45.64	58.25	56.51	67.00	62.02	65.33	52.61	62.65	61.47

Table 5: LAS scores for parsers using predicted morpho-syntactic attributes. First row is LAS using form representation. Rows 2 to 5 are LAS using morphological representation, either one-hot or embedding and either hard decisions or soft decisions.

Appendix A: Supplementary Tables

Table 3 reports results for the predicted attributes experiment. The POS and averaged attributes prediction accuracies are given. Are also reported, scores for the four representation regimes of predicted attributes. Predictions can be either probability distributions (Soft) or argmax (Hard) and either used as such (OH) or passed through an embedding (Emb).

Table 4 reports all the labeled accuracy scores for parsers using either gold lemmas, forms or gold attributes, either as one-hot vectors or as dense embeddings.

Table 5 reports results for the predicted attributes experiment. Are also reported, scores for the four representation regimes of predicted attributes as in table 4. Predictions can be either probability distributions (Soft) or argmax (Hard) and either used as such (OH) or passed through an embedding (Emb).