Tracking Typological Traits of Uralic Languages in Distributed Language Representations

Bjerva, Johannes; Augenstein, Isabelle

Published in:
Proceedings, Fourth International Workshop on Computational Linguistics for Uralic Languages

DOI:
10.18653/v1/W18-02

Publication date:
2018

Document version
Early version, also known as pre-print

Tracking Typological Traits of Uralic Languages in Distributed Language Representations

Johannes Bjerva
Department of Computer Science
University of Copenhagen
Denmark
bjerva@di.ku.dk

Isabelle Augenstein
Department of Computer Science
University of Copenhagen
Denmark
augenstein@di.ku.dk

Abstract

Although linguistic typology has a long history, computational approaches have only recently gained popularity. The use of distributed representations in computational linguistics has also become increasingly popular. A recent development is to learn distributed representations of language, such that typologically similar languages are spatially close to one another. Although empirical successes have been shown for such language representations, they have not been subjected to much typological probing. In this paper, we first look at whether this type of language representations are empirically useful for model transfer between Uralic languages in deep neural networks. We then investigate which typological features are encoded in these representations by attempting to predict features in the *World Atlas of Language Structures*, at various stages of fine-tuning of the representations. We focus on Uralic languages, and find that some typological traits can be automatically inferred with accuracies well above a strong baseline.

Tiivistelmä

Vaikka kielitypologialla on pitkä historia, laskentamenetelmät ovat vasta viime aikoina saavuttaneet suosiota. Myös hajautettujen esitysten käyttö laskennallisessa kielitieteessä on tullut suositummaksi. Viimeaikainen kehitys on hajautetun kieliedustuksen oppiminen, kuten että samanlaiset kielet ovat lähellä toisiaan. Vaikka empiirisiä tuloksia onkin saavutettu, ei niille ole tehty paljoakaan typologista tutkimusta. Tässä artikkelissa tutkitaan ensin, ovatko tämänlaiset kieliedustukset empiirisesti käyttökelpoisia, kun kyseessä on uralilaisten kielten ”model transfer” syvissä neuroverkoissa. Tutkimme myös, mitä typologisia piirteitä voimme löytää kieliedustuksissa, yrittämällä ennustaa ominaisuuksia jotka saamme *World Atlas of Language Structures*:n kautta. Keskitymme uralilaisiin kieliin ja löydämme, että jotkin typologiset piirteet voidaan automaattisesti päätellä selvästi vahvan perustason yläpuolelle.

1 Introduction

For more than two and a half centuries, linguistic typologists have studied languages with respect to their structural and functional properties, thereby implicitly classifying languages as being more or less similar to one another, by virtue of such properties (Haspelmath, 2001; Velupillai, 2012). Although typology has a long history (Herder, 1772; Gabelentz, 1891; Greenberg, 1960, 1974; Dahl, 1985; Comrie, 1989; Haspelmath, 2001; Croft, 2002), computational approaches have only recently gained popularity (Dunn et al., 2011; Wälchli, 2014; Östling, 2015; Bjerva and Börstell, 2016; Deri and Knight, 2016; Cotterell and Eisner, 2017; Peters et al., 2017; Asgari and Schütze, 2017; Malaviya et al., 2017). One part of traditional typological research can be seen as assigning sparse explicit feature vectors to languages, for instance manually encoded in databases such as the World Atlas of Language Structures (WALS, Dryer and Haspelmath, 2013). A recent development which can be seen as analogous to this, is the process of learning distributed language representations in the form of dense real-valued vectors, often referred to as *language embeddings* (Tsvetkov et al., 2016; Östling and Tiedemann, 2017; Malaviya et al., 2017). These language embeddings encode typological properties of language, reminiscent of the sparse features in WALS, or even of parameters in Chomsky's Principles and Parameters framework (Chomsky, 1993; Chomsky and Lasnik, 1993; Chomsky, 2014).

In this paper, we investigate the usefulness

of explicitly modelling similarities between languages in deep neural networks using language embeddings. To do so, we view NLP tasks for multiple Uralic languages as different aspects of the same problem and model them in one model using multilingual transfer in a multi-task learning model. Multilingual models frequently follow a hard parameter sharing regime, where all hidden layers of a neural network are shared between languages, with the language either being implicitly coded in the input string (Johnson et al., 2017), given as a language ID in a one-hot encoding (Ammar et al., 2016), or as a language embedding (Östling and Tiedemann, 2017). In this paper, we both explore multilingual modelling of Uralic languages, and probe the language embeddings obtained from such modelling in order to gain novel insights about typological traits of Uralic languages. We aim to answer the following three research questions (**RQ**s).

RQ 1 To what extent is model transfer between Uralic languages for PoS tagging mutually beneficial?

RQ 2 Are distributed language representations useful for model transfer between Uralic languages?

RQ 3 Can we observe any explicit typological properties encoded in these distributed language representations when considering Uralic languages?

2 Data

2.1 Distributed language representations

There are several methods for obtaining distributed language representations by training a recurrent neural language model (Mikolov et al., 2010) simultaneously for different languages (Tsvetkov et al., 2016; Östling and Tiedemann, 2017). In these recurrent multilingual language models with long short-term memory cells (LSTM, Hochreiter and Schmidhuber, 1997), languages are embedded into a n-dimensional space. In order for multilingual parameter sharing to be successful in this setting, the neural network is encouraged to use the language embeddings to encode features of language. Other work has explored learning language embeddings in the context of neural machine translation (Malaviya et al., 2017). In this work, we explore the embeddings trained by Östling and Tiedemann (2017), both in their original state, and by further tuning them for PoS tagging.

2.2 Part-of-speech tagging

We use PoS annotations from version 2 of the Universal Dependencies (Nivre et al., 2016). We focus on the four Uralic languages present in the UD, namely Finnish (based on the Turku Dependency Treebank, Pyysalo et al., 2015), Estonian (Muischnek et al., 2016), Hungarian (based on the Hungarian Dependency Treebank, Vincze et al., 2010), and North Sámi (Sheyanova and Tyers, 2017). As we are mainly interested in observing the language embeddings, we down-sample all training sets to 1500 sentences (approximate number of sentences in the Hungarian data), so as to minimise any size-based effects.

2.3 Typological data

In the experiments for RQ3, we attempt to predict typological features. We extract the features we aim to predict from WALS (Dryer and Haspelmath, 2013). We consider features which are encoded for all four Uralic languages in our sample.

3 Method and experiments

We approach the task of PoS tagging using a fairly standard bi-directional LSTM architecture, based on Plank et al. (2016). The system is implemented using DyNet (Neubig et al., 2017). We train using the Adam optimisation algorithm (Kingma and Ba, 2014) over a maximum of 10 epochs, using early stopping. We make two modifications to the bi-LSTM architecture of Plank et al. (2016). First of all, we do not use any atomic embedded word representations, but rather use only character-based word representations. This choice was made so as to encourage the model not to rely on language-specific vocabulary. Additionally, we concatenate a pre-trained language embedding to each word representation. That is to say, in the original bi-LSTM formulation of Plank et al. (2016), each word w is represented as $\vec{w} + LSTM_c(w)$, where $\vec{w}$ is an embedded word representation, and $LSTM_c(w)$ is the final states of a character bi-LSTM running over the characters in a word. In our formulation, each word w in language l is represented as $LSTM_c(w) + \vec{l}$, where $LSTM_c(w)$ is defined as before, and $\vec{l}$ is an embedded language representation. We use a

two-layer deep bi-LSTM, with 100 units in each layer. The character embeddings used also have 100 dimensions. We update the language representations, $\vec{l}$, during training. The language representations are 64-dimensional, and are initialised using the language embeddings from Östling and Tiedemann (2017). All PoS tagging results reported are the average of five runs, each with different initialisation seeds, so as to minimise random effects in our results.

3.1 Model transfer between Uralic languages

The aim of these experiments is to provide insight into **RQ 1** and **RQ 2**. We first train a monolingual model for each of the four Uralic languages. This model is then evaluated on all four languages, to investigate how successful model transfer between pairs of languages is. Results are shown in Figure 1. Comparing results within each language shows that transfer between Finnish and Estonian is the most successful. This can be expected considering that these are the two most closely related languages in the sample, as both are Finnic languages. Model transfer both to and from the more distantly related languages Hungarian and North Sámi is less successful. There is little-to-no difference in this monolingual condition with respect to whether or not language embeddings are used. As a baseline, we include transfer results when training on Spanish, which we consider a proxy of a distantly related languages. Transferring from Spanish is significantly worse ($p < 0.05$) than transferring from a Uralic language in all settings.

Next, we train a bilingual model for each Uralic language. Each model is trained on the target language in addition to one other Uralic language. Results are shown in Figure 2. Again, transfer between the two Finnic languages is the most successful. Here we can also observe a strong effect of whether or not language embeddings are incorporated in the neural architecture. Including language embeddings allows for both of the Finnic languages to benefit significantly ($p < 0.05$) from the transfer setting, as compared to the monolingual setting. No significant differences are observed for other language pairs.

3.2 Predicting typological features with language embeddings

Having observed that language embeddings are beneficial for model transfer between Uralic languages, we turn to the typological experiments probing these embeddings. The aim of these experiments is to provide insight into **RQ 3**. We investigate typological features from WALS (Dryer and Haspelmath, 2013), focussing on those which have been encoded for the languages included in the UD.

We first train the same neural network architecture as for the previous experiments on all languages in UD version 2. Observing the language embeddings from various epochs of training permits tracking the typological traits encoded in the distributed language representations as they are fine-tuned. In order to answer the research question, we train a simple linear classifier to predict typological traits based on the embeddings. Concretely, we train a logistic regression model, which takes as input a language embedding $\vec{l}_e$ from a given epoch of training, e, and outputs the typological class a language belongs to (as coded in WALS). When e is 0, this indicates the pre-trained language embeddings as obtained from Östling and Tiedemann (2017). Increasing e indicates the number of epochs of PoS tagging during which the language embedding has been updated. All results are the mean of three-fold cross-validation. We are mainly interested in observing two things: i) Which typological traits do language embeddings encode?; ii) To what extent can we track the changes in these language embeddings over the course of fine-tuning for the task of PoS tagging?.

We train the neural network model over five epochs, and investigate differences of classification accuracies of typological properties as compared to pre-trained embeddings. A baseline reference is also included, which is defined as the most frequently occurring typological trait within each category. In these experiments, we disregard typological categories which are rare in the observed sample (i.e. of which we have one or zero examples). Looking at classification accuracy of WALS features, we can see four emerging patterns:

1. The feature is pre-encoded;
2. The feature is encoded by fine-tuning;
3. The feature is not pre-encoded;
4. The feature encoding is lost by fine-tuning.

One example per category is given in Figure 3. Two features based on word-ordering can be seen as belonging in the categories of features which are either pre-encoded or which become encoded during training. The fine-tuned embeddings do not

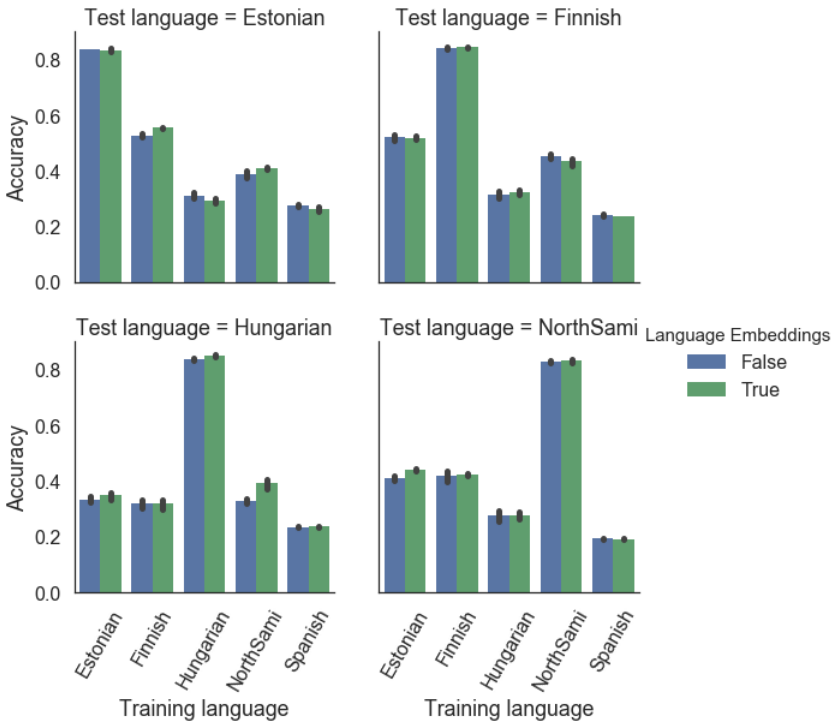

Figure 1: Monolingual PoS training. The x-axes denote the training languages, and the y-axes denote the PoS tagging accuracy on the test language at hand.

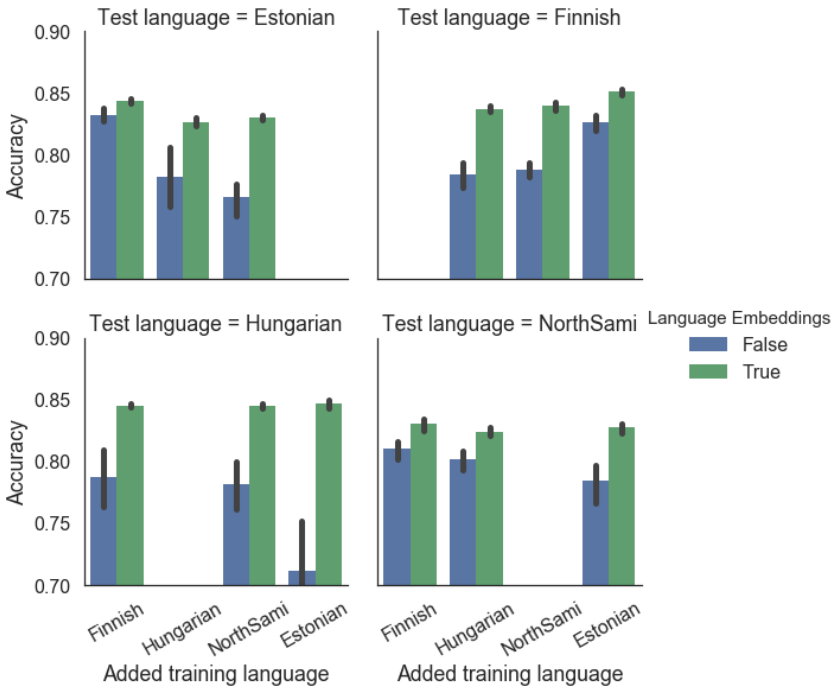

Figure 2: Bilingual PoS training. The x-axes denote the added training languages (in addition to the target language), and the y-axes denote the PoS tagging accuracy on the test language at hand.

encode the feature for whether pronominal subjects are expressed, or the feature for whether a predicate nominal has a zero copula.

3.2.1 Predicting Uralic typological features

Finally, we attempt to predict typological features for the four Uralic languages included in our sample, as shown in Figure 4. Similarly to the larger language sample in Figure 3, the Uralic language embeddings also both gain typological information in some respects, and lose information in other respects. For instance, the pre-trained embeddings are not able to predict ordering of adpositions and noun phrase in the Uralic languages, whereas training on PoS tagging for two epochs adds this information.

4 Discussion

4.1 Language embeddings for Uralic model transfer

In the monolingual transfer setting, we observed that transferring from more closely-related languages was relatively beneficial. This is expected, as the more similar two languages are, the easier it ought to be for the model to directly apply what it learns from one language to the other. Concretely, we observed that transferring between the two Finnic languages in our sample, Finnish and Estonian, worked relatively well. We further observed that including language embeddings in this setting had little-to-no effect on the results. This can be explained by the fact that the language embedding used is the same throughout the training phase, as only one language is used, hence the network likely uses this embedding to a very low extent.

In bilingual settings, omitting the language embeddings results in a severe drop in tagging accuracy in most cases. This is likely because that treating our sample of languages as being the same language introduces a large amount of confusion into the model. This is further corroborated by the fact that treating the two Finnic languages in this manner results in a relatively small drop in accuracy.

Including language embeddings allows for the model transfer setting to be beneficial for the more closely related languages. This bodes well for the low-resource case of many Uralic languages in particular, and possibly for low-resource NLP in general. In the cases of the more distantly related language pairings, including language embeddings does not result in any significant drop in accuracy. This indicates that using language embeddings at least allows for learning a more compact model without any significant losses to performance.

4.2 Language embeddings for Uralic typology

Interestingly, the language embeddings are not only a manner for the neural network to identify which language it is dealing with, but are also used to encode language similarities and typological features. To contrast, the neural network could have learned something akin to a one-hot encoding of each language, in which case the languages could easily have been told apart, but classification of typological features would have been constantly at baseline level.

Another interesting finding is the fact that we can track the typological traits in the distributed language representations as they are fine-tuned for the task at hand. This has the potential to yield insight on two levels, of interest both to the more engineering-oriented NLP community, as well as the more linguistically oriented CL community. A more in-depth analysis of these embeddings can both show what a neural network is learning to model, in particular. Additionally, these embeddings can be used to glean novel insights and answer typological research questions for languages which, e.g., do not have certain features encoded in WALS.

In the specific case of Uralic languages, as considered in this paper, the typological insights we gained are, necessarily, ones that are already known for these languages. This is due to the fact that we simply evaluated our method on the features present for the Uralic languages in WALS. It is nonetheless encouraging for this line of research that we, e.g., could predict WALS feature 86A (*Order of Genitive and Noun*) based solely on these embeddings, and training a very simple classifier on a sample consisting exclusively of non-Uralic languages.

5 Conclusions and future work

We investigated model transfer between the four Uralic languages Finnish, Estonian, Hungarian and North Sámi, in PoS tagging, focussing on the effects of using language embeddings. We

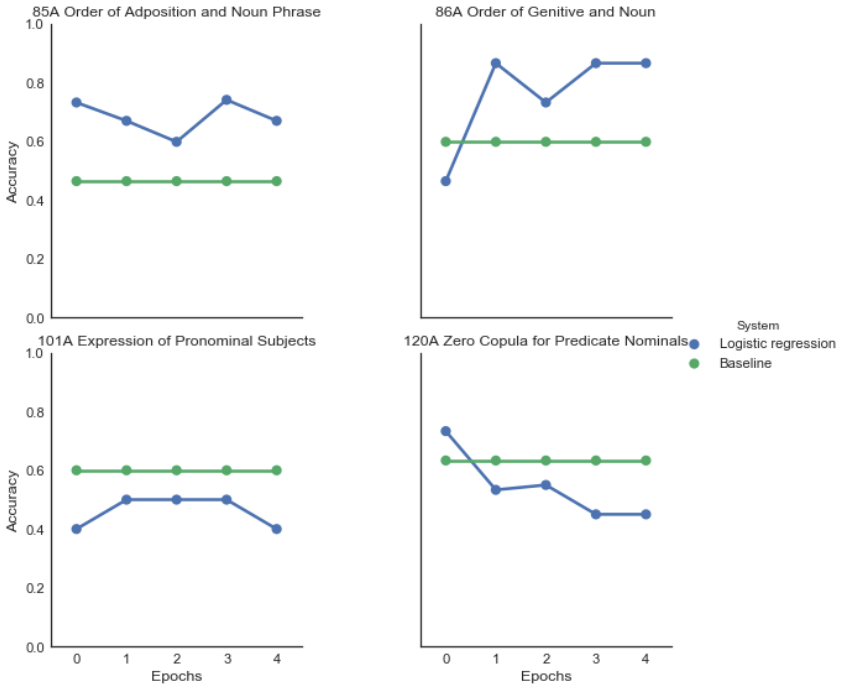

Figure 3: Predicting typological features in WALS. The x-axes denote number of epochs the language embeddings have been fine-tuned for. The y-axe denotes classification accuracy for the typological feature at hand.

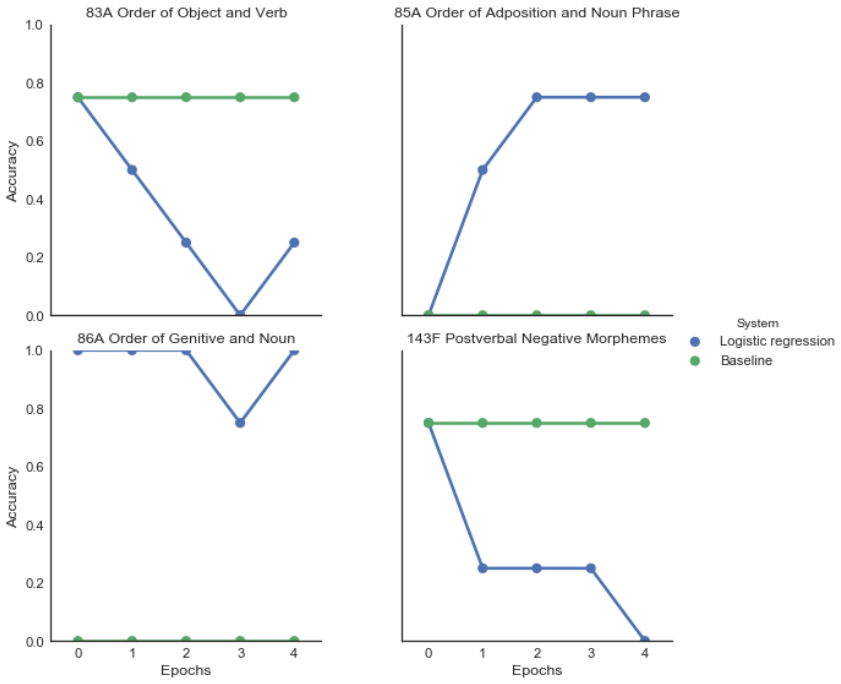

Figure 4: Predicting typological features in Uralic languages. The x-axes denote number of epochs the language embeddings have been fine-tuned for. The y-axes denote classification accuracy for the typological feature at hand.

found that model transfer is successful between these languages, with the main benefits found between the two Finnic languages (Finnish and Estonian), when using language embeddings. We then turned to an investigation of the typological features encoded in the language embeddings, and found that certain features are encoded. Furthermore, we found that the typological features encoded change when fine-tuning the embeddings. In future work, we will look more closely at how the encoding of typological traits in distributed language representations changes depending on the task on which they are trained.

Acknowledgements

The authors would like to thank Iina Alho for help with translating the abstract to Finnish. We would also like to thank Robert Östling for giving us access to pre-trained language embeddings.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association of Computational Linguistics* 4:431–444. https://aclanthology.info/pdf/Q/Q16/Q16-1031.pdf.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *EMNLP*. Association for Computational Linguistics, pages 113–124.

Johannes Bjerva and Carl Börstell. 2016. Morphological Complexity Influences Verb-Object Order in Swedish Sign Language. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. The COLING 2016 Organizing Committee, pages 137–141.

Noam Chomsky. 1993. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

Noam Chomsky. 2014. *The minimalist program*. MIT press.

Noam Chomsky and Howard Lasnik. 1993. The theory of principles and parameters.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *ACL*. Association for Computational Linguistics, pages 1182–1192.

William Croft. 2002. *Typology and universals*. Cambridge University Press.

Östen Dahl. 1985. *Tense and Aspect Systems*. Basil Blackwell Ltd., NewYork.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *ACL*. Association for Computational Linguistics, pages 399–408.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/.

Michael Dunn, Simon J Greenhill, Stephen C Levinson, and Russell D Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345):79–82.

Georg von der Gabelentz. 1891. *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig.

Joseph Greenberg. 1974. *Language typology: A historical and analytic overview*, volume 184. Walter de Gruyter.

Joseph H Greenberg. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics* 26(3):178–194.

Martin Haspelmath. 2001. *Language typology and language universals: An international handbook*, volume 20. Walter de Gruyter.

J. Herder. 1772. *Abhandlung über den Ursprung der Sprache*. Berlin: Christian Friedrich Voß.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics* 5:339–351. http://aclweb.org/anthology/Q17-1024.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *EMNLP*. Association for Computational Linguistics, pages 2519–2525. http://aclweb.org/anthology/D17-1267.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*. volume 2, page 3.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2016. Estonian dependency treebank: from constraint grammar tagset to universal dependencies. In *LREC*.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, et al. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980* .

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. pages 205–211.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *EACL*. Association for Computational Linguistics, pages 644–649. http://aclanthology.coli.uni-saarland.de/pdf/E/E17/E17-2102.pdf.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively multilingual neural grapheme-to-phoneme conversion. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. pages 19–26.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *ACL*. Association for Computational Linguistics, pages 412–418. https://doi.org/10.18653/v1/P16-2067.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal dependencies for finnish. In *NoDaLiDa*. Linköping University Electronic Press, 109, pages 163–172.

Mariya Sheyanova and Francis M. Tyers. 2017. Annotation schemes in north sámi dependency parsing. In *Proceedings of the 3rd International Workshop for Computational Linguistics of Uralic Languages*. pages 66–75.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *NAACL-HLT*. Association for Computational Linguistics, pages 1357–1366. https://doi.org/10.18653/v1/N16-1161.

Viveka Velupillai. 2012. *An introduction to linguistic typology*. John Benjamins Publishing.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *LREC*.

Bernhard Wälchli. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech* 28:355.