# Learning Language Representations for Typology Prediction

**Chaitanya Malaviya** and **Graham Neubig** and **Patrick Littell**
Language Technologies Institute
Carnegie Mellon University
{cmalaviy,gneubig,pwl}@cs.cmu.edu

## Abstract

One central mystery of neural NLP is what neural models "know" about their subject matter. When a neural machine translation system learns to translate from one language to another, does it learn the syntax or semantics of the languages? Can this knowledge be extracted from the system to fill holes in human scientific knowledge? Existing typological databases contain relatively full feature specifications for only a few hundred languages. Exploiting the existence of parallel texts in more than a thousand languages, we build a massive many-to-one neural machine translation (NMT) system from 1017 languages into English, and use this to predict information missing from typological databases. Experiments show that the proposed method is able to infer not only syntactic, but also phonological and phonetic inventory features, and improves over a baseline that has access to information about the languages' geographic and phylogenetic neighbors.[1]

## 1 Introduction

Linguistic typology is the classification of human languages according to syntactic, phonological, and other classes of features, and the investigation of the relationships and correlations between these classes/features. This study has been a scientific pursuit in its own right since the 19th century (Greenberg, 1963; Comrie, 1989; Nichols, 1992), but recently typology has borne practical fruit within various subfields of NLP, particularly on problems involving lower-resource languages.

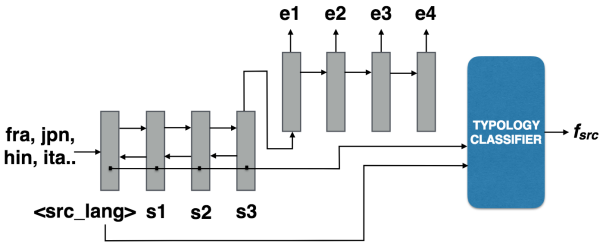


Figure 1: Learning representations from multilingual neural MT for typology classification. (Model MTBOTH)

Typological information from sources like the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), has proven useful in many NLP tasks (O'Horan et al., 2016), such as multilingual dependency parsing (Ammar et al., 2016), generative parsing in low-resource settings (Naseem et al., 2012; Täckström et al., 2013), phonological language modeling and loanword prediction (Tsvetkov et al., 2016), POS-tagging (Zhang et al., 2012), and machine translation (Daiber et al., 2016).

However, the needs of NLP tasks differ in many ways from the needs of scientific typology, and typological databases are often only sparsely populated, by necessity or by design.[2] In NLP, on the other hand, what is important is having a relatively full set of features for the particular group of languages you are working on. This mismatch of needs has motivated various proposals to reconstruct missing entries, in WALS and other databases, from known entries (Daumé III and Campbell, 2007; Daumé III, 2009; Coke et al., 2016; Littell et al., 2017).

In this study, we examine whether we can

[1] Code and learned vectors are available at http://github.com/chaitanyamalaviya/lang-reps

[2] For example, each chapter of WALS aims to provide a statistically balanced set of languages over language families and geographical areas, and so many languages are left out in order to maintain balance.



*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2512–2518
Copenhagen, Denmark, September 7–11, 2017. 

tackle the problem of inferring linguistic typology from parallel corpora, specifically by training a massively multi-lingual neural machine translation (NMT) system and using the learned representations to infer typological features for each language. This is motivated both by prior work in linguistics (Bugarski, 1991; García, 2002) demonstrating strong links between translation studies and tools for contrastive linguistic analysis, work in inferring typology from bilingual data (Östling, 2015) and English as Second Language texts (Berzak et al., 2014), as well as work in NLP (Shi et al., 2016; Kuncoro et al., 2017; Belinkov et al., 2017) showing that syntactic knowledge can be extracted from neural nets on the word-by-word or sentence-by-sentence level. This work presents a more holistic analysis of whether we can discover what neural networks learn about the linguistic concepts of an entire language by aggregating their representations over a large number of the sentences in the language.

We examine several methods for discovering feature vectors for typology prediction, including those learning a language vector specifying the language while training multilingual neural language models (Östling and Tiedemann, 2017) or neural machine translation (Johnson et al., 2016) systems. We further propose a novel method for aggregating the values of the latent state of the encoder neural network to a single vector representing the entire language. We calculate these feature vectors using an NMT model trained on 1017 languages, and use them for typlogy prediction both on their own and in composite with feature vectors from previous work based on the genetic and geographic distance between languages (Littell et al., 2017). Results show that the extracted representations do in fact allow us to learn about the typology of languages, with particular gains for syntactic features like word order and the presence of case markers.

## 2 Dataset and Experimental Setup

**Typology Database:** To perform our analysis, we use the URIEL language typology database (Littell et al., 2017), which is a collection of binary features extracted from multiple typological, phylogenetic, and geographical databases such as WALS (World Atlas of Language Structures) (Collins and Kayne, 2011), PHOIBLE (Moran et al., 2014), Ethnologue (Lewis et al., 2015), and Glottolog (Hammarström et al., 2015). These features are divided into separate classes regarding syntax (e.g. whether a language has prepositions or postpositions), phonology (e.g. whether a language has complex syllabic onset clusters), and phonetic inventory (e.g. whether a language has interdental fricatives). There are 103 syntactical features, 28 phonology features and 158 phonetic inventory features in the database.

**Baseline Feature Vectors:** Several previous methods take advantage of typological implicature, the fact that some typological traits correlate strongly with others, to use known features of a language to help infer other unknown features of the language (Daumé III and Campbell, 2007; Takamura et al., 2016; Coke et al., 2016). As an alternative that does not necessarily require pre-existing knowledge of the typological features in the language at hand, Littell et al. (2017) have proposed a method for inferring typological features directly from the language's $k$ nearest neighbors ($k$-NN) according to geodesic distance (distance on the Earth's surface) and genetic distance (distance according to a phylogenetic family tree). In our experiments, our baseline uses this method by taking the 3-NN for each language according to normalized geodesic+genetic distance, and calculating an average feature vector of these three neighbors.

**Typology Prediction:** To perform prediction, we trained a logistic regression classifier[3] with the baseline $k$-NN feature vectors described above and the proposed NMT feature vectors described in the next section. We train individual classifiers for predicting each typological feature in a class (syntax etc). We performed 10-fold cross-validation over the URIEL database, where we train on 9/10 of the languages to predict 1/10 of the languages for 10 folds over the data.

## 3 Learning Representations for Typology Prediction

In this section we describe three methods for learning representations for typology prediction with multilingual neural models.

**LM Language Vector** Several methods have been proposed to learn multilingual language

[3]We experimented with a non-linear classifier as well, but the logistic regression classifier performed better.

models (LMs) that utilize vector representations of languages (Tsvetkov et al., 2016; Östling and Tiedemann, 2017). Specifically, these models train a recurrent neural network LM (RNNLM; Mikolov et al. (2010)) using long short-term memory (LSTM; Hochreiter and Schmidhuber (1997)) with an additional vector representing the current language as an input. The expectation is that this vector will be able to capture the features of the language and improve LM accuracy. Östling and Tiedemann (2017) noted that, intriguingly, agglomerative clustering of these language vectors results in something that looks roughly like a phylogenetic tree, but stopped short of performing typological inference. We train this vector by appending a special token representing the source language (e.g. "⟨fra⟩" for French) to the beginning of the source sentence, as shown in Fig. 1, then using the word representation learned for this token as a representation of the language. We will call this first set of feature vectors LMVEC, and examine their utility for typology prediction.

**NMT Language Vector** In our second set of feature vectors, MTVEC, we similarly use a language embedding vector, but instead learn a multilingual neural MT model trained to translate from many languages to English, in a similar fashion to Johnson et al. (2016); Ha et al. (2016). In contrast to LMVEC, we hypothesize that the alignments to an identical sentence in English, the model will have a stronger signal allowing it to more accurately learn vectors that reflect the syntactic, phonetic, or semantic consistencies of various languages. This has been demonstrated to some extent in previous work that has used specifically engineered alignment-based models (Lewis and Xia, 2008; Östling, 2015; Coke et al., 2016), and we examine whether these results apply to neural network feature extractors and expand beyond word order and syntax to other types of typology as well.

**NMT Encoder Mean Cell States** Finally, we propose a new vector representation of a language (MTCELL) that has not been investigated in previous work: the average hidden cell state of the encoder LSTM for all sentences in the language. Inspired by previous work that has noted that the hidden cells of LSTMs can automatically capture salient and interpretable information such as syntax (Karpathy et al., 2015; Shi et al., 2016) or sentiment (Radford et al., 2017), we expect that the cell states will represent features that may be linked to the typology of the language. To create vectors for each language using LSTM hidden states, we obtain the mean of cell states ($\boldsymbol{c}$ in the standard LSTM equations) for all time steps of all sentences in each language.[4]

| | Syntax | | Phonology | | Inventory | |
|---|---|---|---|---|---|---|
| | -Aux | +Aux | -Aux | +Aux | -Aux | +Aux |
| NONE | 69.91 | 83.07 | 77.92 | 86.59 | 85.17 | 90.68 |
| LMVEC | 71.32 | 82.94 | 80.80 | 86.74 | 87.51 | 89.94 |
| MTVEC | 74.90 | 83.31 | 82.41 | 87.64 | 89.62 | 90.94 |
| MTCELL | 75.91 | 85.14 | 84.33 | 88.80 | 90.01 | 90.85 |
| MTBOTH | **77.11** | **86.33** | **85.77** | **89.04** | **90.06** | **91.03** |

Table 1: Accuracy of syntactic, phonological, and inventory features using LM language vectors (LMVEC), MT language vectors (MTVEC), MT encoder cell averages (MTCELL) or both MT feature vectors (MTBOTH). Aux indicates auxiliary information of geodesic/genetic nearest neighbors; "NONE -Aux" is the majority class chance rate, while "NONE +Aux" is a 3-NN classification.

## 4 Experiments

### 4.1 Multilingual Data and Training Regimen

To train a multilingual neural machine translation system, we used a corpus of Bible translations that was obtained by scraping a massive online Bible database at `bible.com`.[5] This corpus contains data for 1017 languages. After preprocessing the corpus, we obtained a training set of 20.6 million sentences over all languages.

The implementation of both the LM and NMT models described in §3 was done in the DyNet toolkit (Neubig et al., 2017). In order to obtain a manageable shared vocabulary for all languages, we divided the data into subwords using joint byte-pair encoding of all languages (Sennrich et al., 2016) with 32K merge operations. We

[4] We also tried using the mean of final hidden cell states of the encoder LSTM, but the mean cell state over all words in the sentence gave improved performance. Additionally, we tried using the hidden states $\boldsymbol{h}$, but we found that these had significantly less information and lesser variance, due to being modulated by the output gate at each time step.

[5] A possible concern is that Bible translations may use archaic language not representative of modern usage. However, an inspection of the data did not turn up such archaisms, likely because the bulk of world Bible translation was done in the late 19th and 20th centuries. In addition, languages that do have antique Bibles are also those with many other Bible translations, so the effect of the archaisms is likely limited.

used LSTM cells in a single recurrent layer with 512-dimensional hidden state and input embedding size. The Adam optimizer was used with a learning rate of 0.001 and a dropout of 0.5 was enforced during training.

### 4.2 Results and Discussion

The results of the experiments can be found in Tab. 1. First, focusing on the "-Aux" results, we can see that all feature vectors obtained by the neural models improve over the chance rate, demonstrating that indeed it is possible to extract information about linguistic typology from unsupervised neural models. Comparing LMVEC to MTVEC, we can see a convincing improvement of 2-3% across the board, indicating that the use of bilingual information does indeed provide a stronger signal, allowing the network to extract more salient features. Next, we can see that MTCELL further outperforms MTVEC, indicating that the proposed method of investigating the hidden cell dynamics is more effective than using a statically learned language vector. Finally, combining both feature vectors as MTBOTH leads to further improvements. To measure statistical significance of the results, we performed a paired bootstrap test to measure the gain between NONE+AUX and MTBOTH+AUX and found that the gains for syntax and inventory were significant (p=0.05), but phonology was not, perhaps because the number of phonological features was fewer than the other classes (only 28).

When further using the geodesic/genetic distance neighbor feature vectors, we can see that these trends largely hold although gains are much smaller, indicating that the proposed method is still useful in the case where we have a-priori knowledge about the environment in which the language exists. It should be noted, however, that the gains of LMVEC evaporate, indicating that access to aligned data may be essential when inferring the typology of a new language. We also noted that the accuracies of certain features decreased from NONE-AUX to MTBOTH-AUX, particularly gender markers, case suffix and negative affix, but these decreases were to a lesser extent in magnitude than the improvements.

Interestingly, and in contrast to previous methods for inferring typology from raw text, which have been specifically designed for inducing word order or other syntactic features (Lewis and Xia, 2008; Östling, 2015; Coke et al., 2016), our proposed method is also able to infer information about phonological or phonetic inventory features. This may seem surprising or even counter-intuitive, but a look at the most-improved phonology/inventory features (Tab. 2) shows a number of features in which languages with the "non-default" option (e.g. having uvular consonants or initial velar nasals, *not* having lateral consonants, etc.) are concentrated in particular geographical regions. For example, uvular consonants are not common world-wide, but are common in particular geographic regions like the North American Pacific Northwest and the Caucasus (Maddieson, 2013b), while initial velar nasals are common in Southeast Asia (Anderson, 2013), and lateral consonants are *uncommon* in the Amazon Basin (Maddieson, 2013a). Since these are also regions with a particular and sometimes distinct syntactic character, we think the model may be find-

| Feature | NONE | MT | Gain |
|---|---|---|---|
| S_NUMERAL_AFTER_NOUN | 37.40 | 81.26 | 43.86 |
| S_NUMERAL_BEFORE_NOUN | 46.49 | 83.22 | 36.73 |
| S_POSSESSOR_AFTER_NOUN | 42.05 | 75.60 | 33.55 |
| S_OBJECT_BEFORE_VERB | 50.97 | 80.89 | 29.92 |
| S_ADPOSITION_AFTER_NOUN | 52.41 | 79.10 | 26.69 |
| P_UVULAR_CONTINUANTS | 77.57 | 97.37 | 19.80 |
| P_LATERALS | 67.30 | 86.48 | 19.18 |
| P_LATERAL_L | 64.05 | 78.16 | 14.10 |
| P_LABIAL_VELARS | 82.16 | 95.93 | 13.76 |
| P_VELAR_NASAL_INITIAL | 72.14 | 85.82 | 13.68 |
| I_VELAR_NASAL | 39.89 | 62.08 | 22.20 |
| I_ALVEOLAR_LATERAL_APPROXIMANT | 60.92 | 79.32 | 18.40 |
| I_ALVEOLAR_NASAL | 81.49 | 92.98 | 11.48 |
| I_VOICED_LABIODENTAL_FRICATIVE | 65.75 | 77.10 | 11.36 |
| I_VOICELESS_PALATAL_FRICATIVE | 82.41 | 93.66 | 11.25 |

Table 2: Top 5 improvements from "NONE -Aux" to "MTBOTH -Aux" in the syntax ("S_"), phonology ("P_"), and inventory ("I_") classes.

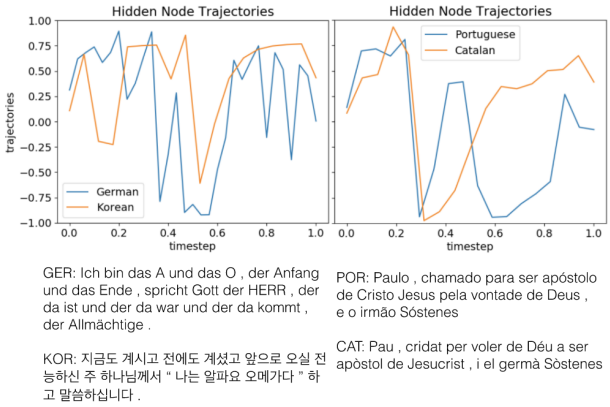


Figure 2: Cell trajectories for sentences in languages where S_OBJ_BEFORE_VERB is either active or inactive.

ing regional clusters through syntax, and seeing an improvement in regionally-distinctive phonology/inventory features as a side effect.

Finally, given that MTCELL uses the feature vectors of the latent cell state to predict typology, it is of interest to observe how these latent cells behave for typologically different languages. In Fig. 2 we examine the node that contributed most to the prediction of "S_OBJ_BEFORE_VERB" (the node with maximum weight in the classifier) for German and Korean, where the feature is active, and Portuguese and Catalan, where the feature is inactive. We can see that the node trajectories closely track each other (particularly at the beginning of the sentence) for Portuguese and Catalan, and in general the languages where objects precede verbs have higher average values, which would be expressed by our mean cell state features. The similar trends for languages that share the value for a typological feature (S_OBJ_BEFORE_VERB) indicate that information stored in the selected hidden node is consistent across languages with similar structures.

## 5 Conclusion and Future Work

Through this study, we have shown that neural models can learn a range of linguistic concepts, and may be used to impute missing features in typological databases. In particular, we have demonstrated the utility of learning representations with parallel text, and results hinted at the importance of modeling the dynamics of the representation as models process sentences. We hope that this study will encourage additional use of typological features in downstream NLP tasks, and inspire further techniques for missing knowledge prediction in under-documented languages.

## Acknowledgments

We thank Lori Levin and David Mortensen for their useful comments and also thank the reviewers for their feedback about this work.

## References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Gregory D.S. Anderson. 2013. The velar nasal. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *Eighteenth Conference on Computational Natural Language Learning (CoNLL)*.

Ranko Bugarski. 1991. Contrastive analysis of terminology and the terminology of contrastive analysis. *Languages in Contact und Contrast. Essays in Contact Linguistics/Edited by Vladimir Ivir and Damir Kalogjera.–Berlin*, pages 73–82.

Reed Coke, Ben King, and Dragomir Radev. 2016. Classifying syntactic regularities for hundreds of languages. *arXiv preprint arXiv:1603.08016*.

Chris Collins and Richard Kayne. 2011. *Syntactic Structures of the World's Languages*. New York University, New York.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. Blackwell, Oxford.

Joachim Daiber, Miloš Stanojević, and Khalil Sima'an. 2016. Universal reordering via linguistic typology. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3167–3176, Osaka, Japan. The COLING 2016 Organizing Committee.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601, Boulder, Colorado. Association for Computational Linguistics.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

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

Noelia Ramón García. 2002. Contrastive linguistics and translation studies interconnected: The corpus-based approach. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, (1).

Joseph Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg, editor, *Universals of Language*, pages 110–113. MIT Press, London.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. *Glottolog 2.6*. Max Planck Institute for the Science of Human History, Jena.

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, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith. 2017. What do recurrent neural network grammars learn about syntax? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1249–1258, Valencia, Spain. Association for Computational Linguistics.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. *Ethnologue: Languages of the World, Eighteenth edition*. SIL International, Dallas, Texas.

William D Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing, Volume II*, pages 685–690.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Ian Maddieson. 2013a. Lateral consonants. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ian Maddieson. 2013b. Uvular consonants. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

Steven Moran, Daniel McCloy, and Richard Wright. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics.

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*.

Joanna Nichols. 1992. *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago.

Helen O'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

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 *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia. Association for Computational Linguistics.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

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 *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal POS tagset. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1368–1378. Association for Computational Linguistics.