



From Phonology to Syntax

Unsupervised Linguistic Typology at Different Levels with Language Embeddings

Bjerva, Johannes; Augenstein, Isabelle

Published in:

Proceedings, 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

Publication date:

2018

Document version

Early version, also known as pre-print

Citation for published version (APA):

Bjerva, J., & Augenstein, I. (2018). From Phonology to Syntax: Unsupervised Linguistic Typology at Different Levels with Language Embeddings. In *Proceedings, 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies : (Long Papers)* (Vol. 1, pp. 907-916). Association for Computational Linguistics.

From Phonology to Syntax: Unsupervised Linguistic Typology at Different Levels with Language Embeddings

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

A core part of linguistic typology is the classification of languages according to linguistic properties, such as those detailed in the World Atlas of Language Structure (WALS). Doing this manually is prohibitively time-consuming, which is in part evidenced by the fact that only 100 out of over 7,000 languages spoken in the world are fully covered in WALS.

We learn distributed language representations, which can be used to predict typological properties on a massively multilingual scale. Additionally, quantitative and qualitative analyses of these language embeddings can tell us how language similarities are encoded in NLP models for tasks at different typological levels. The representations are learned in an unsupervised manner alongside tasks at three typological levels: phonology (grapheme-to-phoneme prediction, and phoneme reconstruction), morphology (morphological inflection), and syntax (part-of-speech tagging).

We consider more than 800 languages and find significant differences in the language representations encoded, depending on the target task. For instance, although Norwegian Bokmål and Danish are typologically close to one another, they are phonologically distant, which is reflected in their language embeddings growing relatively distant in a phonological task. We are also able to predict typological features in WALS with high accuracies, even for unseen language families.

1 Introduction

For more than two and a half centuries, linguistic typologists have studied languages with respect to their structural and functional 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; Cotterell and Eisner, 2017; Asgari and Schütze, 2017; Malaviya et al., 2017; Bjerva and Augenstein, 2018). 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). We hypothesise that these language embeddings encode typological properties of language, reminiscent of the features in WALS, or even of parameters in Chomsky’s Principles and Parameters framework (Chomsky, 1993).

In this paper, we model languages in deep neural networks using language embeddings, considering three typological levels: phonology, morphology and syntax. We consider four NLP tasks to be representative of these levels: grapheme-to-phoneme prediction and phoneme reconstruction, morphological inflection, and part-of-speech tagging. We pose three research questions (RQs):

RQ 1 Which typological properties are encoded in task-specific distributed language representations, and can we predict phonological, morphological and syntactical properties of languages using such representations?

RQ 2 To what extent do the encoded properties change as the representations are fine tuned for tasks at different linguistic levels?

RQ 3 How are language similarities encoded in fine-tuned language embeddings?

One of our key findings is that language representations differ considerably depending on the target task. For instance, for grapheme-to-phoneme mapping, the differences between the representations for Norwegian Bokmål and Danish increase rapidly during training. This is due to the fact that, although the languages are typologically close to one another, they are phonologically distant.

2 Related work

Computational linguistics approaches to typology are now possible on a larger scale than ever before due to advances in neural computational models. Even so, recent work only deals with fragments of typology compared to what we consider here.

Computational typology has to a large extent focused on exploiting word or morpheme alignments on the massively parallel New Testament, in approximately 1,000 languages, in order to infer word-order (Östling, 2015) or assign linguistic categories (Asgari and Schütze, 2017). Wälchli (2014) similarly extracts lexical and grammatical markers using New Testament data. Other work has taken a computational perspective on language evolution (Dunn et al., 2011), and phonology (Cotterell and Eisner, 2017; Alishahi et al., 2017).

Language embeddings In this paper, we follow an approach which has seen some attention in the past year, namely the use of distributed language representations, or *language embeddings*. Some typological experiments are carried out by Östling and Tiedemann (2017), who learn language embeddings via multilingual language modelling and show that they can be used to reconstruct genealogical trees. Malaviya et al. (2017) learn language embeddings via neural machine translation, and predict word-ordering features.

Contributions Our work bears the most resemblance to Bjerva and Augenstein (2018), who fine-tune language embeddings on the task of PoS tagging, and investigate how a handful of typological properties are coded in these for four Uralic languages. We expand on this and thus contribute to previous work by: (i) introducing novel qualitative investigations of language embeddings, in addition to thorough quantitative evaluations; (ii) considering four tasks at three different typological levels; (iii) considering a far larger sample of several hundred languages; and (iv) grounding the language representations in linguistic theory.

3 Background

3.1 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. In this paper, we explore the embeddings trained by Östling and Tiedemann (2017), both in their original state, and by further tuning them for our four tasks. While other work has looked at the types of representations encoded in different layers of deep neural models (Kádár et al., 2017), we choose to look at the representations only in the bottom-most embedding layer. This is motivated by the fact that we look at several tasks using different neural architectures, and want to ensure comparability between these.

3.1.1 Language embeddings as continuous Chomskyan parameter vectors

We now turn to the theoretical motivation of the language representations. The field of NLP is littered with distributional word representations, which are theoretically justified by distributional semantics (Harris, 1954; Firth, 1957), summarised in the catchy phrase *You shall know a word by the company it keeps* (Firth, 1957). We argue that language embeddings, or distributed representations of language, can also be theoretically motivated by Chomsky’s Principles and Parameters framework (Chomsky and Lasnik, 1993; Chomsky, 1993, 2014). Language embeddings encode languages as dense real-valued vectors, in which the dimensions are reminiscent of the parameters found in this framework. Briefly put, Chomsky argues that languages can be described in terms of principles (abstract rules) and parameters (switches) which can be turned either on or off for a given language (Chomsky and Lasnik, 1993). An example of such a switch might represent the positioning of the head of a clause (i.e. either head-initial or head-final). For English, this switch

would be set to the ‘initial’ state, whereas for Japanese it would be set to the ‘final’ state. Each dimension in an n -dimensional language embedding might also describe such a switch, albeit in a more continuous fashion. The number of dimensions used in our language representations, 64, is a plausible number of parameter vector dimensions (Dunn et al., 2011). If we were able to predict typological features using such representations, this lends support to the argument that languages, at the very least, *can* be represented by theoretically motivated parameter vectors, with the given dimensionality.

3.2 Typological features in the World Atlas of Language Structure

In the experiments for **RQ1** and **RQ2** we predict typological features extracted from WALS (Dryer and Haspelmath, 2013). We choose to investigate three linguistic levels of language: phonology, morphology, and syntax. This is motivated by three factors: (i) these features are related to NLP tasks for which data is available for a large language sample; (ii) the levels cover a range from basic phonological and morphological structure, to syntactic structure, allowing us to approach our research question from several angles; and (iii) the features in these categories are coded in WALS for a relatively large selection of languages. We extract the three feature sets which represent these typological levels of language from WALS.¹

Phonological features cover 20 features ranging from descriptions of the consonant and vowel inventories of a particular language to presence of tone and stress markers. As an example, consider WALS feature 13A (Tone).² This feature takes three feature values: (i) *no tones*, (ii) *simple tone system*, and (iii) *complex tone system*. Most Indo-European languages, such as English, Spanish, and Russian, do not have any tones (i). Norwegian and Swedish are exceptions to this, as they both have simple tone systems (ii) similar to that in Japanese. Finally, complex tone systems (iii) are typically found in several African languages as well as languages in South-East Asia.

Morphological features cover a total of 41 features. We consider the features included in the Morphology chapter as well as those included in

the Nominal Categories chapter to be morphological in nature.³ This includes features such as the number of genders, usage of definite and indefinite articles and reduplication. As an example, consider WALS feature 37A (Definite Articles).⁴ This feature takes five values: (i) *Definite word distinct from demonstrative*, (ii) *Demonstrative word used as definite article*, (iii) *Definite affix*, (iv) *No definite, but indefinite article*, and (v) *No definite or indefinite article*. Again, most Indo-European languages fall into category (i), with Norwegian, Swedish, and Danish as relative outliers in category (iii).

Word-order features cover 56 features, encoding properties such as the ordering of subjects, objects and verbs. As an example, consider WALS feature 81A (Order of Subject, Object and Verb).⁵ This feature takes all possible combinations of the three word classes as its feature values, with the addition of a special class for *No dominant order*. Most languages in WALS fall into the categories SOV (41.0%) and SVO (35.4%).

4 Method

The general set-up of the experiments in this paper are as follows. We aim at answering our three research questions dealing with typological properties and similarities as encoded in language embeddings. In order to do this, we attempt to predict typological features as they are encoded in WALS, using language embeddings which have been fine-tuned during training on tasks related to different typological properties. The main interest in this paper is therefore not on how well each model performs on a given NLP task, but rather on what the resulting language embeddings encode.

Concretely, we use language embeddings \vec{l}_i from a given training iteration of a given task as input to a k-NN classifier, which outputs the typological class a language belongs to (as coded in WALS). We train separate classifiers for each typological property and each target task. When $i = 0$, this indicates the pre-trained language embeddings as obtained from Östling and Tiedemann (2017). Increasing i indicates the number of iterations over which the system at hand has been

¹These are defined in the chapter structure in WALS: <http://wals.info/chapter>

²<http://wals.info/feature/13A>

³This choice was made as, e.g., feature 37A (Definite Articles) includes as a feature value whether a definite affix is used.

⁴<http://wals.info/feature/37A>

⁵<http://wals.info/feature/81A>

trained. In each experiment, for a given iteration i , we consider each $\vec{l}_i \in L$ where L is the intersection $L_{task} \cap L_{pre}$, where L_{task} is the set of languages for a given task, and L_{pre} is the set of the languages for which we have pre-trained embeddings.

All results in the following sections are the mean of three-fold cross-validation, and the mean over the WALS features in each given category.⁶ We run the experiments in a total of three settings: (i) evaluating on randomly selected language/feature pairs from a task-related feature set; (ii) evaluating on an *unseen language family from a task-related feature set*; (iii) evaluating on randomly selected language/feature pairs from all WALS feature sets. This allows us to establish how well we can predict task-related features given a random sample of languages (i), and a sample from which a whole language family has been omitted (ii). Finally, (iii) allows us to compare the task-specific feature encoding with a general one.

A baseline reference is also included, which is defined as the most frequently occurring typological trait within each category.⁷ For instance, in the morphological experiments, we only consider the 41 WALS features associated with the categories of morphology and nominal categories. The overlap between languages for which we have data for morphological inflection and languages for which these WALS features are coded is relatively small (fewer than 20 languages per feature). This small dataset size is why we have opted for a non-parametric k -Nearest Neighbours classifier for the typological experiments. We use $k = 1$, as several of the features take a large number of class values, and might only have a single instance represented in the training set.

Table 1 shows the datasets we consider (detailed in later sections), the typological class they are related to, the size of the language sample in the task, and the size of the intersection $L_{task} \cap L_{pre}$. The number of pre-trained language embeddings, $|L_{pre}|$, is 975 in all cases. We focus the evaluation for each task-specific language embedding set on the typological property relevant to that dataset. In

⁶The mean accuracy score is a harsh metric, as some features are very difficult to predict due to them, e.g., being very language specific or taking a large number of different values.

⁷The languages represented in several of the tasks under consideration have a high Indo-European bias. Hence, several of the properties are relatively uniformly distributed, providing us with a strong baseline.

addition, we also evaluate on a set of all typological properties in WALS. Note that the evaluation on all properties is only comparable to the evaluation on each specific property, as the set of languages under consideration differs between tasks.

Dataset	Class	$ L_{task} $	$ L_{task} \cap L_{pre} $
G2P	Phonology	311	102
ASJP	Phonology	4,664	824
SIGMORPHON	Morphology	52	29
UD	Syntax	50	27

Table 1: Overview of tasks and datasets.

5 Phonology

5.1 Grapheme-to-phoneme

We use grapheme-to-phoneme (G2P) as a proxy of a phonological task (Deri and Knight, 2016; Peters et al., 2017). The dataset contains over 650,000 such training instances, for a total of 311 languages (Deri and Knight, 2016). The task is to produce a phonological form of a word, given its orthographic form and the language in which it is written. Crucially, this mapping is highly different depending on the language at hand. For instance, take the word written *variation*, which exists in both English and French:

(English, variation) \rightarrow ,vɛəri'eɪʃən
(French, variation) \rightarrow ,vɑʁja'sjɔ̃

5.1.1 Experiments and Analysis

We train a sequence-to-sequence model with attention for the task of grapheme-to-phoneme mapping.⁸ The model takes as input the characters of each source form together with the language embedding for the language at hand and outputs a predicted phonological form. Input and output alphabets are shared across all languages. The system is trained over 3,000 iterations.

Quantitative results Since we consider Grapheme-to-Phoneme as a phonological task, we focus the quantitative evaluation on phonological features from WALS. We run experiments using the language embeddings as features for a simple k -NN classifier. The results in Table 2 indicate that G2P is a poor proxy for language phonology, however, as typological properties pertaining to phonology are not encoded. That is to say, the k -NN results do not outperform the baseline, and

⁸The system is described in detail in Section 8.

performance is on par even after fine tuning (no significant difference, $p > 0.05$). In the unseen setting, however, we find that pre-trained language embeddings are significantly better ($p > 0.05$) at predicting the phonological features than both fine-tuned ones and the baseline.

System / features	Random phon.	Unseen phon.	All feat.
Most Frequent Class	*75.46%	65.57%	79.90%
k-NN (pre-trained)	71.45%	*86.54%	80.39%
k-NN (fine-tuned)	71.66%	82.36%	79.17%

Table 2: Accuracies on prediction of WALS features with language embeddings fine-tuned on Grapheme-to-Phoneme mapping. Asterisks indicate results significantly higher than both other conditions ($p < 0.05$).

Qualitative results We now turn to why this task is not a good proxy of phonology. The task of grapheme-to-phoneme is more related to the processes in the diachronic development of the writing system of a language than it is to the actual genealogy or phonology of the language. This is evident when considering the Scandinavian languages Norwegian and Danish which are typologically closely related, and have almost exactly the same orthography. In spite of this fact, the phonology of the two languages differs drastically due to changes in Danish phonology, which impacts the mapping from graphemes to phonemes severely. Hence, the written forms of the two languages should be very similar, which makes the language embeddings based on language modelling highly similar to one another. However, when the embeddings are fine-tuned on a task taking orthography as well as phonology into account, this is no longer the case. Figure 1 shows that the language

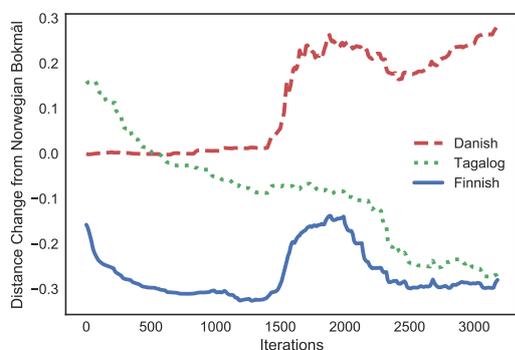


Figure 1: Language similarities between Norwegian, and Danish/Tagalog/Finnish, as G2P-based embeddings are fine-tuned.

embeddings of Norwegian Bokmål and Danish diverge from each other, which is especially striking when comparing to the converging with the typologically much more distant languages Tagalog and Finnish which. However, the absolute difference between Norwegian Bokmål and both Tagalog/Finnish is still greater than that of Norwegian Bokmål and Danish even after 3,000 iterations.

5.2 Phonological reconstruction

As a second phonological task, we look at phonological reconstruction using word lists from the Automated Similarity Judgement Program (ASJP, Wichmann et al. (2016)). This resource contains word lists of at least 40 words per language for more than 4,500 languages. The task we consider is to reproduce a given source phonological form, also given the language, for instance:

(English, wat3r) \rightarrow wat3r

The intuition behind these experiments is that languages with similar phonetic inventories will be grouped together, as reflected in changes in the language embeddings.

5.2.1 Experiments and Analysis

We train a sequence-to-sequence model with attention, framed as an auto-encoding problem, using the same sequence-to-sequence architecture and setup as for the grapheme-to-phoneme task. The model takes as input the characters of each source form together with the language embedding for the language at hand and outputs the predicted target form which is identical to the source form.

Quantitative results Since we also consider phonological reconstruction to be a phonological task, we focus the quantitative evaluation on phonological features from WALS. As with the G2P experiments, Table 3 shows that the fine-tuned embeddings do not offer predictive power above the most frequent class baseline ($p < 0.01$). Observing the changes in the language embeddings reveals that the embeddings are updated to a very small extent, indicating that these are not used by the model to a large extent. This can be explained by the fact that the task is highly similar for each language, and that the model largely only needs to learn to copy the input string.

We do, however find that evaluating on a set with an unseen language family does yield results

significantly above baseline levels with the pre-trained embeddings ($p > 0.01$), which together with the G2P results indicate that the language modelling objective does encode features to some extent related to phonology.

System / features	Random phon.	Unseen phon.	All feat.
Most Frequent Class	*59.39%	63.71%	*58.12%
k-NN (pre-trained)	53.02%	*77.44%	51.6%
k-NN (fine-tuned)	53.09%	*77.45%	51.9%

Table 3: Accuracies on prediction of WALS features with language embeddings fine-tuned on Phonological Reconstruction. Asterisks indicate results significantly higher than non-bold conditions ($p < 0.01$).

6 Morphology

6.1 Morphological inflection

We use data from the Unimorph project, specifically the data released for the SIGMORPHON-2017 shared task (Cotterell et al., 2017).⁹ This data covers 52 languages, thereby representing a relatively large typological variety. Whereas the shared task has two subtasks, namely inflection and paradigm cell filling, we only train our system using the inflection task. This was a choice of convenience, as we are not interested in solving the task of morphological paradigm cell filling, but rather observing the language embeddings as they are fine-tuned. Furthermore we focus on the high-resource setting in which we have access to 10,000 training examples per language. The inflection subtask is to generate a target inflected form given a lemma with its part-of-speech as in the following example:

```
(release, V;V.PTCP;PRS) -> releasing
```

6.1.1 Morphological experiments

We train a sequence-to-sequence model with attention over 600 iterations, using the same sequence-to-sequence architecture from the previous tasks.

Quantitative results Since this is considered a morphological task, Table 4 contains results using the language embeddings to predict morphological properties. The fine-tuned language embeddings in this condition are able to predict morphological properties in WALS significantly above baseline levels and pre-trained embeddings ($p < 0.01$). We further also obtain significantly better results in the

⁹<https://unimorph.github.io/>

unseen setting ($p < 0.01$), in which the language family evaluated on is not used in training. This indicates that these properties are important to the model when learning the task at hand.

System / Features	Random morph.	Unseen morph.	All feat.
Most Frequent Class	77.98%	85.68%	84.12%
k-NN (pre-trained)	74.49%	88.83%	84.97%
k-NN (fine-tuned)	*82.91%	*91.92%	84.95%

Table 4: Accuracies on prediction of WALS features with language embeddings fine-tuned on morphological inflection. Asterisks indicate results significantly higher than both other conditions ($p < 0.01$).

Qualitative results The performance of the fine-tuned embeddings on prediction of morphological features is above baseline for most features. For 18 out of the 35 features under consideration both the baseline and k-NN performances are at 100% from the outset, so these are not considered here.¹⁰ Figure 2 shows two of the remaining 17 features.¹¹ We can observe two main patterns: For some features such as 49A (Number of cases), fine-tuning on morphological inflection increases the degree to which the features are encoded in the language embeddings. This can be explained by the fact that the number of cases in a language is central to how morphological inflection is treated by the model. For instance, languages with the same number of cases might benefit from sharing certain parameters. On the other hand, the feature 38A (Indefinite Articles) mainly encodes whether the indefinite word is distinct or not from the word for *one*, and it is therefore not surprising that this is not learned in morphological inflection.

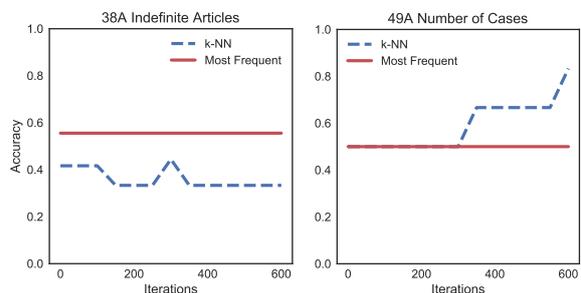


Figure 2: Prediction of two morphological features in WALS with morphological language embeddings, one data point per 50 iterations.

¹⁰This is partially explained by the fact that certain categories were completely uniform in the small sample as well as by the Indo-European bias in the sample.

¹¹The full feature set is included in the Supplements.

7 Word order

7.1 Part-of-speech tagging

We use PoS annotations from version 2 of the Universal Dependencies (Nivre et al., 2016). As we are mainly interested in observing the language embeddings, we down-sample all training sets to 1,500 sentences (approximate number of sentences of the smallest data sets) so as to minimise any size-related effects.

7.1.1 Word-order experiments

We approach the task of PoS tagging using a fairly standard bi-directional LSTM architecture based on Plank et al. (2016), detailed in Section 8.

Quantitative results Table 5 contains results on WALS feature prediction using language embeddings fine-tuned on PoS tagging. We consider both the set of word order features, which are relevant for the dataset, and a set of all WALS features. Using the fine-tuned embeddings is significantly better than both the baseline and the pre-trained embeddings ($p < 0.05$), in both the random and the unseen conditions, indicating that the model learns something about word order typology. This can be expected, as word order features are highly relevant when assigning a PoS tag to a word.

System / features	Random W-Order	Unseen W-Order	All feat.
Most Frequent Class	76.81%	82.47%	82.93%
k-NN (pre-trained)	76.66%	92.76%	82.69%
k-NN (fine-tuned)	*80.81%	*94.48%	83.55%

Table 5: Accuracies on prediction of WALS features with language embeddings fine-tuned on PoS tagging. Asterisks indicate the result in bold significantly outperforming both other conditions ($p < 0.05$).

Qualitative results We now turn to the syntactic similarities between languages as encoded in the fine-tuned language embeddings. We consider a set of the North-Germanic languages Icelandic, Swedish, Norwegian Nynorsk, Danish, Norwegian Bokmål, the West-Germanic language English, and the Romance languages Spanish, French, and Italian. We apply hierarchical clustering using UPGMA (Michener and Sokal, 1957) to the pre-trained language embeddings of these languages.¹² Striking here is that English is grouped together with the Romance languages. This can be explained by the fact that English has a large amount of vocabulary stemming from Romance

¹²Included in the Supplements due to space restrictions.

loan words, which under the task of language modelling results in a higher similarity with such languages. We then cluster the embeddings fine-tuned on PoS tagging in the same way. In this condition, English has joined the rest of the Germanic languages’ cluster. This can be explained by the fact that, in terms of word ordering and morpho-syntax, English is more similar to these languages than it is to the Romance ones.

We can also observe that, whereas the orthographically highly similar Norwegian Bokmål and Danish form the first sub-cluster in the pre-trained condition, Norwegian Nynorsk replaces Danish in this pairing when fine-tuning on PoS tagging. This can be explained by the fact that morpho-syntactic similarities between the two written varieties of Norwegian are more similar to one another.

8 Implementation

8.1 Sequence-to-sequence modelling

The system architecture used in the sequence-to-sequence tasks, i.e., G2P, phonological reconstruction, and morphological inflection is depicted in Figure 3. The system is based on that developed by Östling and Bjerva (2017) and is implemented using Chainer (Tokui et al., 2015). We modify the architecture by concatenating a language embedding, \vec{l} , to the character embeddings before encoding. In the grapheme-to-phoneme and phonological reconstruction experiments, the one-hot feature mapping before decoding is irrelevant and therefore omitted. The rest of the hyperparameters are the same as in Östling and Bjerva (2017).

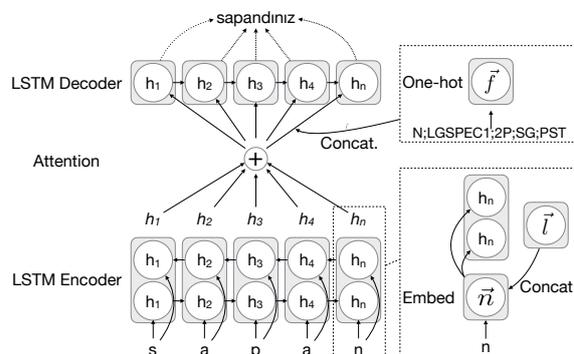


Figure 3: System architecture used in the seq-to-seq tasks (morphological inflection, G2P, and phonological reconstruction). Figure adapted with permission from Östling and Bjerva (2017).

8.2 Sequence labelling

This system is based on Plank et al. (2016), and 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. In our formulation, each word w is represented as $LSTM_c(w) + \vec{l}$, where $LSTM_c(w)$ is the final states of a character bi-LSTM running over the characters in a word and \vec{l} is an embedded language representation. We use a two-layer deep bi-LSTM with 100 units in each layer, and 100-dimensional character embeddings. The rest of the hyper-parameters are the same as in Plank et al. (2016).¹³

9 Discussion and Conclusions

The language embeddings obtained by fine-tuning on linguistic tasks at various typological levels were found to include typological information somehow related to the task at hand. This lends some support to the theoretical foundations of such representations, in that it shows that it is possible to learn something akin to a vector of continuous Chomskyan parameters (Chomsky, 1993).

9.1 RQ1: Encoding of typological features in task-specific language embeddings

The features which are encoded depend to a large degree on the task at hand. The language embeddings resulting from the phonological tasks did not encode phonological properties in the sense of WALS features, whereas the pre-trained ones did. The morphological language embeddings were found to encode morphological features, and the PoS language embeddings were similarly found to encode word order features.

A promising result is the fact that we were able to predict typological features for unseen language families. That is to say, without showing, e.g., a single Austronesian training instance to the k-NN

¹³Both modified systems are included in the Supplements, and will be made publicly available.

classifier, typological features could still be predicted with high accuracies. This indicates that we can predict typological features with language embeddings, even for languages for which we have no prior typological knowledge.

Table 6 contains a comparison of the top five and bottom five feature prediction accuracies for the ASJP task.¹⁴ In the case of the phonologically oriented ASJP task it is evident that the embeddings still encode something related to phonology, as four out of five best features are phonological.

9.2 RQ2: Change in encoding of typological features

The changes in the features encoded in language embeddings are relatively monotonic. Features are either learnt, forgotten, or remain static throughout training. This indicates that the language representations converge towards a single optimum.

9.3 RQ3: Language similarities

Training language embeddings in the task of multilingual language modelling has been found to reproduce trees which are relatively close matches to more traditional genealogical trees (Östling and Tiedemann, 2017). We show a similar analysis considering pre-trained and PoS fine-tuned embeddings, and it is noteworthy that fine-tuning on PoS tagging in this case yielded a tree more faithful to genealogical trees, such as those represented on glottolog.org.

Another striking result in terms of language similarities in fine-tuned language embedding spaces was found in the G2P task. We here found that the phonological differences between some otherwise similar languages, such as Norwegian Bokmål and Danish, were accurately encoded.

Task	Feature	WALS Chapter	Accuracy
ASJP	6A	Phonology	89.45
	18A	Phonology	88.91
	20A	Morphology	82.74
	19A	Phonology	82.58
	7A	Phonology	80.97
	144A	Word Order	14.48
	144L	Word Order	10.07
	62A	Nominal Syntax	10.00
	81B	Word Order	9.52
	133A	Lexicon	8.18

Table 6: Top 5 and bottom 5 accuracies in feature prediction using phonological language embeddings.

¹⁴See the Supplement for the remaining tasks.

References

- Afra Alishahi, Marie Barking, and Grzegorz Chrupała. 2017. Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, pages 368–378.
- 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 Isabelle Augenstein. 2018. Tracking Typological Traits of Uralic Languages in Distributed Language Representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics for Uralic Languages (IWCLUL)*.
- 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.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, et al. 2017. Conllsigmorphon 2017 shared task: Universal morphological reinflection in 52 languages. *arXiv preprint arXiv:1706.09031*.
- William Croft. 2002. *Typology and universals*. Cambridge University Press.
- Östen Dahl. 1985. *Tense and Aspect Systems*. Basil Blackwell Ltd., New York.
- 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.
- J. R. Firth. 1957. A synopsis of linguistic theory pages 1930–1955. 1952–1959:1–32.
- 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.
- Z. Harris. 1954. Distributional structure. *Word* 10:146–162.
- 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.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural network. *Computational Linguistics* 43:761–780.
- Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *EMNLP*. Association for Computational Linguistics, pages 2519–2525.
- Charles D Michener and Robert R Sokal. 1957. A quantitative approach to a problem in classification. *Evolution* 11(2):130–162.
- Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Inter-speech*. volume 2, page 3.
- 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 Johannes Bjerva. 2017. Su-rug at the conll-sigmorphon 2017 shared task: Morphological inflection with attentional sequence-to-sequence models. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*. Association for Computational Linguistics, pages 110–113.
- Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *EACL*. Association for Computational Linguistics, pages 644–649.
- 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.
- Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.
- Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqi, 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.
- Viveka Velupillai. 2012. *An introduction to linguistic typology*. John Benjamins Publishing.
- 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.
- Søren Wichmann, Eric W. Holman, and Cecil H. Brown (eds.). 2016. The ASJP Database (version 17).