What Do Language Representations Really Represent?

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

Robert Östling
Stockholm University
Department of Linguistics

Maria Han Veiga
University of Zurich
Institute of Computational Science

Jörg Tiedemann
University of Helsinki
Department of Digital Humanities

Isabelle Augenstein
University of Copenhagen
Department of Computer Science

A neural language model trained on a text corpus can be used to induce distributed representations of words, such that similar words end up with similar representations. If the corpus is multilingual, the same model can be used to learn distributed representations of languages, such that similar languages end up with similar representations. We show that this holds even when the multilingual corpus has been translated into English, by picking up the faint signal left by the source languages. However, just as it is a thorny problem to separate semantic from syntactic similarity in word representations, it is not obvious what type of similarity is captured by language representations. We investigate correlations and causal relationships between language representations learned from translations on one hand, and genetic, geographical, and several levels of structural similarity between languages on the other. Of these, structural similarity is found to correlate most strongly with language representation similarity, whereas genetic relationships—a convenient benchmark used for evaluation in previous work—appears to be a confounding factor. Apart from implications about translation effects, we see this more generally as a case where NLP and linguistic typology can interact and benefit one another.

Submission received: 12 July 2018; revised version received: 22 October 2018; accepted for publication: 13 December 2018.

doi:10.1162/COLI_a_00351

1. Introduction

Words can be represented with distributed word representations, currently often in the form of word embeddings. Similarly to how words can be embedded, so can languages, by associating each language with a real-valued vector known as a **language representation**, which can be used to measure similarities between languages. This type of representation can be obtained by, for example, training a multilingual model for some NLP task (Johnson et al. 2017; Malaviya, Neubig, and Littell 2017; Östling and Tiedemann 2017). The focus of this work is on the evaluation of similarities between such representations. This is an important area of work, as computational approaches to typology (Dunn et al. 2011; Cotterell and Eisner 2017; Bjerva and Augenstein 2018) have the potential to answer research questions on a much larger scale than traditional typological research (Haspelmath 2001). Furthermore, having knowledge about the relationships between languages can help in NLP applications (Ammar et al. 2016), and having incorrect interpretations can be detrimental to multilingual NLP efforts. For instance, if the similarities between languages in an embedded language space were to be found to encode geographical distances (Figure 1), any conclusions drawn from use of these representations would not likely be of much use for most NLP tasks. The importance of having deeper knowledge of what such representations encapsulate is further hinted at by both experiments with interpolation of language vectors (Östling and Tiedemann 2017), as well as multilingual translation models (Johnson et al. 2017).

Several previous authors have done preliminary investigations into the structure of language representations: Östling and Tiedemann (2017), Malaviya, Neubig, and Littell (2017), and Johnson et al. (2017) in the context of language modeling and machine translation, all of them using multilingual data. In this work we follow up on the findings of Rabinovich, Ordan, and Wintner (2017), who, by using language representations consisting of manually specified feature vectors, find that the structure of a language representation space is approximately preserved by translation. However, their analysis only stretches as far as finding a correlation between their language representations and genetic distance, even though the latter is correlated to several other factors. We apply a multilingual language model to this problem, and evaluate the learned representations against a set of three language properties: (i) genetic distance (families), (ii) a novel measure of syntactic similarity (structural), and (iii) distance of language communities (geographical). We investigate:

RQ1. In what way do different language representations encode language similarities? In particular, is genetic similarity what is really captured?

RQ2. What causal relations can we find between language representation similarities?

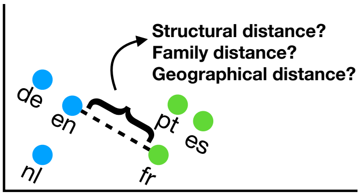
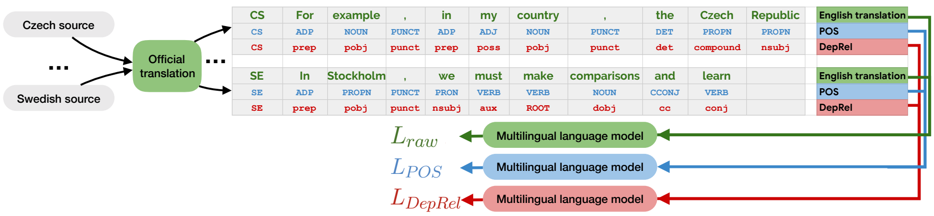


Figure 1
Language representations in a two-dimensional space. What do their similarities represent?



CS	For	example	,	in	my	country	,	the	Czech	Republic
CS	ADP	NOUN	PUNCT	ADP	ADJ	NOUN	PUNCT	DET	PROPN	PROPN
CS	prep	pobj	punct	prep	poss	pobj	punct	det	compound	nsubj
SE	In	Stockholm	,	we	must	make	comparisons	and	learn	
SE	ADP	PROPN	PUNCT	PRON	VERB	VERB	NOUN	CCONJ	VERB	
SE	prep	pobj	punct	nsubj	aux	ROOT	dobj	cc	conj	

Figure 2
Problem illustration. Given official translations from EU languages to English, we train multilingual language models on various levels of abstractions, encoding the source languages. The resulting source language representations (L_{raw}, etc.) are evaluated.

1.1 Contributions

Our work is most closely related to Rabinovich, Ordan, and Wintner (2017), who investigate representation learning on monolingual English sentences, which are translations from various source languages to English from the Europarl corpus (Koehn 2005). They use a feature-engineering approach to predict source languages and learn an Indo-European family tree using their language representations, showing that there are significant traces of the source languages in translations. They use features based on sequences of part-of-speech (POS) tags, function words, and cohesive markers. Additionally, they posit that the similarities found between their representations encode the genetic relationships between languages. We show that this is not the strongest explanation of the similarities, as a novel syntactic measure offers far more explanatory value, which we further substantiate by investigating causal relationships between language representations and similarities (Pearl 2009). This is an important finding as it highlights the need for thoroughly substantiating linguistic claims made based on empirical findings. Further, understanding what similarities are encoded in language embeddings provides insights into how language embeddings could be used for downstream multilingual NLP tasks. If language representations are used for transfer learning to low-resource languages, having an incorrect view of the structure of the language representation space can be dangerous. For instance, the standard assumption of genetic similarity would imply that the representation of the Gagauz language (Turkic, spoken mainly in Moldova) should be interpolated from the genetically very close Turkish, but this would likely lead to poor performance in syntactic tasks because the two languages have diverged radically in syntax relatively recently.

2. Method

Figure 2 illustrates the data and problem we consider in this paper. We are given a set of English gold-standard translations from the official languages of the European Union, based on speeches from the European Parliament.[1] We wish to learn language representations based on these data, and investigate the linguistic relationships that hold between the resulting representations (**RQ1**). It is important to abstract away from the surface forms of the translations as, for example, speakers from certain regions

[1] This is the exact same data as used by Rabinovich, Ordan, and Wintner (2017), originating from Europarl (Koehn 2005).

will tend to talk about the same issues, or places. We therefore introduce three levels of abstraction: (i) training on function words and POS; (ii) training on only POS tags (POS in Figure 2); (iii) training on sequences of dependency relation tags (DepRel in Figure 2), and constituent tags. This annotation is obtained using UDPipe (Straka, Hajic, and Straková 2016).

2.1 Language Representations

For each level of abstraction, we train a multilingual neural language model in order to obtain representations (vectors in $\mathbb{R}^k$) that we can analyze further (**RQ1**). Note that this model is multilingual in the sense that we model the *source language* of each input sequence, whereas the input sequences themselves are, for example, sequences of POS tags. Our model is a multilingual language model using a standard two-layer long short-term memory architecture. Multilinguality is approached similarly to Östling and Tiedemann (2017), who include a language representation at each time-step. That is to say, each input is represented both by a symbol representation, c, and a language representation, $l \in L$. Because the set of language representations L is updated during training, the resulting representations encode linguistic properties of the languages. Whereas Östling and Tiedemann (2017) model hundreds of languages, we model only English—however, we redefine L to be the set of source languages from which our translations originate.

3. Family Trees from Translations

We now consider the language representations obtained from training our neural language model on the input sequences with different representations of the text (characters, POS sequences, etc.). We cluster the language representations—vectors in $\mathbb{R}^k$—hierarchically[2] and compute similarities between our generated trees and the gold tree of Serva and Petroni (2008), using the distance metric from Rabinovich, Ordan, and Wintner (2017).[3] Our generated trees yield comparable results to previous work (Table 1).

Language Modeling using Lexical Information and POS Tags. Our first experiments deal with training directly on the raw translated texts. This is likely to bias representations by speakers from different countries talking about specific issues or places (as in Figure 2), and gives the model comparatively little information to work with as there is no explicit syntactic information available. As a consequence of the lack of explicit syntactic information, it is unsurprising that the results (**LM-Raw** in Table 1) only marginally outperform the random baseline.

To abstract away from the content and negate the geographical effect we train a new model on only function words and POS. This performs almost on par with LM-Raw (**LM-Func** in Table 1), indicating that the level of abstraction reached is not sufficient to capture similarities between languages. We next investigate whether we can successfully abstract away from the content by removing function words, and only using POS tags (**LM-POS** in Table 1). Although Rabinovich, Ordan, and Wintner (2017)

[2] Following Rabinovich, Ordan, and Wintner (2017), we use the same implementation of Ward's algorithm. We use vector cosine distance rather than Euclidean distance because it is more natural for language vector representations, where the vector magnitude is not important.

[3] Trees not depicted here can be found in the supplements: `http://dx.doi.org/10.1162/coli_a_00351`.

Table 1
Tree distance evaluation (lower is better, cf. §5.1).

Condition	Mean	St.d.
Raw text (LM-Raw)	0.527	-
Function words and POS (LM-Func)	0.556	-
Only POS (LM-POS)	0.517	-
Phrase-structure (LM-Phrase)	0.361	-
Dependency Relations (LM-Deprel)	**0.321**	-
POS trigrams (ROW17)	0.353	0.06
Random (ROW17)	0.724	0.07

produce sensible trees by using trigrams of POS and function words, we do not obtain such trees in our most similar settings. One hypothesis for why this is the case is the differing architectures used—indicating that our neural architecture does not pick up on the trigram-level statistics present in their explicit feature representations.

Language Modeling on Phrase Structure Trees and Dependency Relations. To force the language model to predict as much syntactic information as possible, we train on bracketed phrase structure trees. Note that this is similar to the target side of Vinyals et al. (2015). All content words are replaced by POS tags, and function words are kept. This results in a vocabulary of 289 items (phrase and POS tags and function words). Syntactic information captures more relevant information for reconstructing trees than previous settings (**LM-Phrase** in Table 1), yielding trees of similar quality to previous work.

We also compare to the Universal Dependencies (UD) formalism, as we train the language model on tuples encoding the dependency relation and POS tag of a word, the head direction, and the head POS tag (**LM-Deprel** in Table 1). The **LM-Phrase** and **LM-Deprel** models yield the best results overall, due to their having access to higher levels of abstraction via syntax. The fact that sufficient cues for the source languages can be found here shows that source language affects the grammatical constructions used (cf. Gellestam 1986).

4. Comparing Languages

Our main contribution is to investigate whether genetic distance between languages is captured by language representations, or if other distance measures provide more explanation (**RQ1**). Having shown that our language representations can reproduce genetic trees on par with previous work, we now compare the language embeddings using three different types of language distance measures: *genetic distance* estimated by methods from historical linguistics, *geographical distance* of speaker communities, and a novel measure for the *structural distances* between languages.

4.1 Genetic Distance

Following Rabinovich, Ordan, and Wintner (2017), we use phylogenetic trees from Serva and Petroni (2008) as our gold-standard representation of genetic distance (Figure 3). For meaningful and fair comparison, we also use the same distance metric.

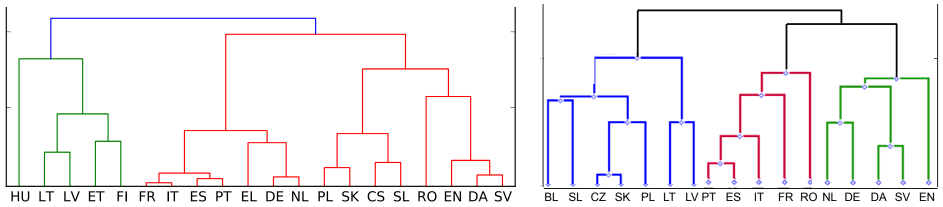


Figure 3
Clustering based on dependency link statistics from UD (left), and the genetic tree from Serva and Petroni (2008) (right). Which type of similarity do language representations really represent?

The metric considers a tree of N leaves, l_n. The weighted distance between two leaves in a tree τ, denoted $D_\tau(l_n, l_m)$, is the sum of the weights of all edges on the shortest path between these leaves. The distance between a generated tree, g, and the gold tree, τ, can then be calculated by summing the square of the differences between all leaf-pair distances (Rabinovich, Ordan, and Wintner 2017):

$$Dist(\tau, g) = \sum_{n,m \in N} (D_\tau(l_n, l_m) - D_g(l_n, l_m))^2$$

4.2 Geographical Distance

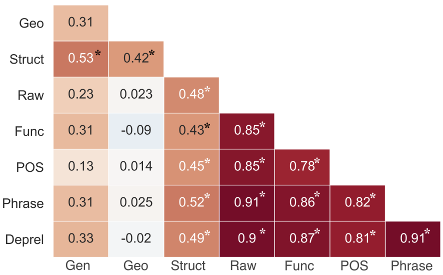
We rely on the coordinates provided by Glottolog (Hammarström, Forkel, and Haspelmath 2017). These are by necessity approximate, because the geography of a language cannot accurately be reduced to a single point denoting the geographical center point of where its speakers live. Still, this provides a way of testing the influence of geographical factors such as language contact or political factors affecting the education system.

4.3 Structural Distance

To summarize the structural properties of each language, we use counts of dependency links from the UD treebanks, version 2.1 (Nivre et al. 2017). Specifically, we represent each link by combining head and dependent POS, dependency type, and direction. This yields 8,607 combinations, so we represent each language by a 8,607-dimensional normalized vector, and compute the cosine distance between these language representations.

Figure 3 shows the result of clustering these vectors (Ward clustering, cosine distance). Although strongly correlated with genealogical distance, significant differences can be observed. Romanian, as a member of the Balkan sprachbund, is distinct from the other Romance languages. The North Germanic (Danish, Swedish) and West Germanic (Dutch, German) branches are separated through considerable structural differences, with English grouped with the North Germanic languages despite its West Germanic origin. The Baltic languages (Latvian, Lithuanian) are grouped with the nearby Finnic languages (Estonian, Finnish) rather than their distant Slavic relatives.

This idea has been explored previously by Chen and Gerdes (2017), who use a combination of relative frequency, length, and direction of deprels. We, by comparison, achieve an even richer representation by also taking head and dependent POS into account.



	Gen	Geo	Struct	Raw	Func	POS	Phrase
Geo	0.31						
Struct	0.53*	0.42*					
Raw	0.23	0.023	0.48*				
Func	0.31	-0.09	0.43*	0.85*			
POS	0.13	0.014	0.45*	0.85*	0.78*		
Phrase	0.31	0.025	0.52*	0.91*	0.86*	0.82*	
Deprel	0.33	-0.02	0.49*	0.9 *	0.87*	0.81*	0.91*

Figure 4
Correlations between similarities (Genetic, Geo., and Struct.) and language representations (Raw, Func, POS, Phrase, Deprel). Significance at $p < 0.001$ is indicated by *.

5. Analysis of Similarities

Although we are able to reconstruct phylogenetic language trees in a similar manner to previous work, we wish to investigate whether genetic relationships between languages really is what our language representations represent.

We generate distance matrices A_ρ, where each entry $a_{i,j}$ represents the ρ-similarity between the i^{th} and j^{th} languages, using the three similarity measures outlined in §4. Then, the entries in A_{gen} contain pairwise genetic distances, computed by summing the weights of all edges on the shortest path between two leaves (languages). Similarly, the entries in A_{geo} contain the geographical distance between countries associated with the languages. Lastly, the entries in A_{struct} contain the cosine distance between the language representations, which are encoded in 8,607-dimensional normalized vectors.

Figure 4 shows the Spearman correlation coefficients between each pair of these matrices. The strongest correlations can be found between the language embeddings, showing that they have similar representations. The correlations between our three distance measures are also considerable (e.g., between geographical and structural distances). This is expected, as languages that are close to one another geographically tend to be similar due to language contact, and potentially shared origins (Velupillai 2012).

What Do Language Representations Really Represent? Most interestingly, the language embedding similarities correlate the most strongly with the structural similarities, rather than the genetic similarities, thus answering **RQ1**. Although previous work by Rabinovich, Ordan, and Wintner (2017) has shown that relatively faithful phylogenetic trees can be reconstructed, we have found an alternative interpretation to these results with much stronger similarities to structural similarities. This indicates that, as often is the case, although similarities between two factors can be found, this is not necessarily the factor with the highest explanatory value (Roberts and Winters 2013).

6. Causal Inference

We further strengthen our analysis by investigating **RQ2**, looking at the relationships between our variables in a Causal Network (Pearl 2009). We use a variant of the Inductive Causation algorithm, namely, IC* (Verma and Pearl 1992). It takes a distribution as input, and outputs a partially directed graph that denotes the (potentially) causal

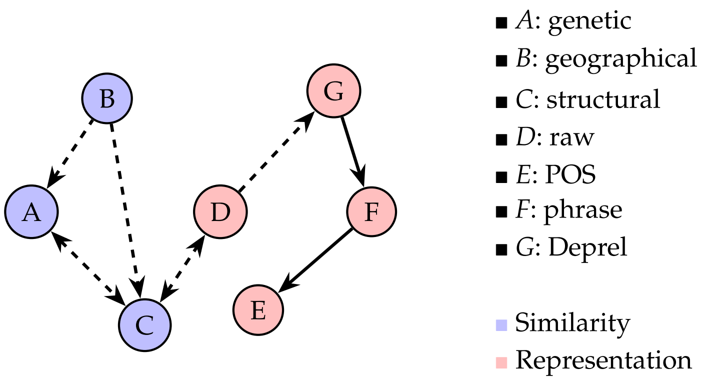


Figure 5
Causal network generated by IC*.

relationships found between each node in the graph. Here, the nodes represent our similarity measures and language embedding distances. The edges in the resulting graph can denote genuine causation (unidirectional edges), potential causation (dashed unidirectional edges), spurious associations (bidirectional edges), and undetermined relationships (undirected edges) (Pearl 2009). Running the algorithm on our distribution based on all the distance measures and language embeddings from this work yields a graph with the following properties, as visualized in Figure 5.[4]

We observe two clusters, marking associations between distance measures and language representations. Interestingly, the only link found between the clusters is an association between the structural similarities and our raw model. This further strengthens our argument, as the fact that no link is found to the genetic similarities shows that our alternative explanation has higher explanatory value, and highlights the need for controlling for more than a single linguistic factor when seeking explanations for one’s results.

7. Discussion and Conclusions

We train language representations on three levels of syntactic abstraction, and explore three different explanations to what language representations represent: genetic, geographical, and structural distances. On the one hand, we extend on previous work by showing that phylogenetic trees can be reconstructed using a variety of language representations (Rabinovich, Ordan, and Wintner 2017). On the other, contrary to a claim of Rabinovich, Ordan, and Wintner (2017), we show that structural similarities between languages are a better predictor of language representation similarities than genetic similarities. As interest in computational typology is increasing in the NLP community (Östling 2015; Bjerva and Augenstein 2018; Gerz et al. 2018; Ponti et al. 2018), we advocate for the necessity of explaining typological findings through comparison.

[4] The IC* algorithm uses pairwise correlations to find sets of conditional independencies between variables at $p < 0.001$, and constructs a minimal partially directed graph that is consistent with the data.

References

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *TACL*, 4:431–444.

Bjerva, Johannes and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *NAACL-HLT*.

Chen, Xinying and Kim Gerdes. 2017. Classifying languages by dependency structure. Typologies of delexicalized universal dependency treebanks. In *DepLing*, pages 54–63.

Cotterell, Ryan and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *ACL*.

Dunn, Michael, 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.

Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. *Translation Studies in Scandinavia*, 1:88–95.

Gerz, Daniela, Ivan Vulic, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *EMNLP*.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2017. Glottolog 3.0. *Jena: Max Planck Institute for the Science of Human History.* (Available online at http://glottolog.org, accessed on 2017-05-15.).

Haspelmath, Martin. 2001. *Language Typology and Language Universals: An International Handbook*, volume 20, Walter de Gruyter.

Johnson, Melvin, 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. *TACL*, 5:339–351.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*.

Malaviya, Chaitanya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *EMNLP*, pages 2519–2525.

Nivre, Joakim, et al. 2017. Universal Dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University.

Östling, Robert. 2015. Word order typology through multilingual word alignment. In *ACL-IJCNLP*, pages 205–211.

Östling, Robert and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *EACL*.

Pearl, Judea. 2009. *Causality*, Cambridge University Press.

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *arXiv preprint arXiv:1807.00914*.

Rabinovich, Ella, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *ACL*.

Roberts, Seán and James Winters. 2013. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PloS One*, 8(8):e70902.

Serva, Maurizio and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EPL*, 81(6):68005.

Straka, Milan, Jan Hajic, and Jana Straková. 2016. UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *LREC*.

Velupillai, Viveka. 2012. *An Introduction to Linguistic Typology*, John Benjamins Publishing.

Verma, Thomas and Judea Pearl. 1992. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the 8th Conference on Uncertainty and Artificial Intelligence*, pages 323–330.

Vinyals, Oriol, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *NIPS*.