Measuring Lexical Semantic Variation using Word Embeddings[1]

Damir Cavar, Indiana University

1 Introduction

This paper discusses an approach to an unsupervised study of lexical semantic variation across languages, dialects, and linguistic variants that is based on a comparison of Distributed Semantics models of lexical items. To achieve this I am using word vectors and embeddings trained on large corpora. My focus in this article is on the South-Slavic languages and variants, Bosnian, Croatian, and Serbian, and taking into account text, corpora, and language models that are explicitly written in Serbo-Croatian. Our focus here is to quantify the lexical overlap and the semantic fields or properties of the lexical items, using a purely unsupervised empirical study based on language use data.

There is a long history of studies related to similarities and dissimilarities between the languages of the Balkans, which I will ignore here entirely. The notion of language as a potentially defining feature for

1. This work has been presented together with Dejan Ivković at the Balkan Conference in Billings, Montana, in May 2018. I am grateful for Dejan's help and comments on this work. We are both grateful for all the inspiring comments from the conference participants. None of this would be possible without the support from Gisbert Fanselow during my time in Potsdam and and long time after.

some concept like state, nation, or ethnicity can deviate from the linguists' notion of what determines a language. In this study my focus is not on *language* as anything else but the communication device for groups of people, manifested in some form in the language faculty of the speaker community and every individual. I assume that language is a dynamic object that is subject to dynamic changes driven by the behavior of the speaker community and the individuals' language faculty. For the study here, the main interest is in the methodology of studying dynamics of the lexicon and semantic properties exploiting distributional lexical models derived using simple neural network architectures and large text corpora.

A statement from Firth (1968) that can be found in many recent papers on distributional semantics summarizes the general idea: "You shall know a word by the company it keeps." This is of course an oversimplification that can only be seen as a rough observational tendency that we can exploit in a more general empiricist approach to lexical semantics that has some practical value. We do not assume this kind of descriptive and purely distributional model to be sufficient in a theoretical approach to lexical semantics.

Vector Space models, using geometrical measures, have been suggested as a tool for the analysis of lexical properties, for example, by Charniak (1997). An overview of the approaches before the Deep Learning and neural network wave can be found in Baroni & Lenci (2010), Turney & Pantel (2010), Erk (2012), and Clark (2014). To capture distributional properties of words in corpora, one could, for example, count the number of times certain words occur in the context of a specific lexical item. This would be represented in form of a vector of scalars that contain the counts or the relative frequency as an estimate of the probability of other words occurring in the context of a particular lexical item.

With the dawn of Deep Learning and the new connectionist wave over the last decade, word embeddings emerged as an elegant way to encode lexical items in form of numerical vectors, capturing aspects of their semantic dimensions that are derived from their distributional properties.

In the following we will discuss geometry-based approaches to lexical semantics and new word embedding approaches used in neural network modeling, before we turn to our analysis of South-Slavic languages using

such embedding models.

2 Approaching Models of Word Meaning with Geometry

In a bag-of-word (BoW) approach, one would assume that the entire corpus is represented in a table with the number of rows and columns corresponding to the number of different words in the corpus (types). Each row could represent one target word and each column—a context word for the target word. The resulting table or matrix represents word-to-word model for a specific corpus. The numbers in the cells correspond to the counts or the number of times a specific type (column-word) occurs in the local context of *n*-words left and right of a target or row-word in a specific corpus. One can experiment with the size of the context for enumeration or BoW selection. We could chose *n* to be 5 words left and right of a target word, or experiment with different window sizes for different analyzes or purposes in our modeling. We then analyze the corpus and count how many times a certain word occurs in the context-window of size *n* of our target word. Consider Table 1.1, where the type *big* occurred 3 times in a specific context-window of the word *dog*.

	...	*the*	*a*	*big*	...
dog	...	45	32	3	...
cat	...	39	27	1	...
walked	...	67	49	1	...

Table 1.1: Bag-of-Words in Context Counts

For a large corpus, the number of rows and columns in a table like 1.1 could have columns and rows determined by the vocabulary size. If a language has a rich morphology, this table could be even larger and the counts would be lower. There are various ways to optimize and improve the quantitative property. As mentioned, one can experiment with the window size for the BoW approach. We could also remove certain types of words, for example stop words, if we were interested only in content words that would play a larger role in the semantic properties of the

target word. Nevertheless, the size of this kind of matrix could be quite large, depending on the corpus size and the text genre.

The model in Table 1.1 captures only general local context. Linguistic properties, however, are often directional in the sense that semantic and syntactic selection or modification is typically to or from one specific direction in some language. In our case, when looking at some South-Slavic languages, modifiers like adjectives would preferably occur to the left of a head noun, or a complement clause would occur to the right of its governing verb. We expect the context position left or right of target words to be significant for distributional models and to capture semantic properties much better.

To capture contextual distributional properties using the BoW approach, we could generate frequency vectors that represent the left and the right context, respectively, as in Table 1.2.

target	**left**					**right**				
	…	**the**	**a**	**big**	…	…	**the**	**a**	**big**	…
dog	…	45	32	3	…	…	0	0	0	…
cat	…	39	27	1	…	…	0	0	0	…
walked	…	0	0	0	…	…	67	49	1	…

Table 1.2: Left and right combined BoW context vectors

It is quite obvious, when looking at Table 1.2, that such a directional BoW vectorization approach for distributional word properties in terms of geometry captures much better distinct lexical semantic properties.

The disadvantage of such an approach is that we would double the dimensionality of the word vectors, thus also the computation and memory requirements for any kind of computational approach based on geometry as an expression of lexical similarities. The bigger problem with larger vector sizes is that of sparseness in the dimensions, with many dimensions being 0. We could generate a more detailed vector model by taking precise positions into account as for example by tracking for every context word whether it occurred one, two, or more words to the left or right of the target word. This would expand our model even further and lead to even more sparseness in the models.

2.1 Similarity Metrics

The common approaches to measure word similarities would be based on Euclidean Distance or Cosine Similarity. In Euclidean Distance we measure the absolute distance between two points *p* and *q* in *n*-dimensional space by taking the square root of the sum of the squares of the distance for each coordinate, as in equation 1.1.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1.1}$$

While Euclidean Distance is appropriate for normalized vectors, that is, if the vectors are based on word frequencies from one corpus, it would not be appropriate for comparisons of word frequencies from different corpora, if the corpora are significantly different in size. A normalized measure of similarity for such vectors would be the Cosine Similarity as in 1.2, which measures the angle between two vectors, by dividing the dot-product between by the product of the magnitude of each vector, thus ignoring the relative length of each vector.

$$cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1.2}$$

Independent of the method that we discussed above, by representing distributional properties of words in form of numerical vectors we can express the intuition that more similar words are closer to each other in terms of any of the geometrical distance metrics, i.e. Euclidean Distance or Cosine Similarity.

2.2 Dense Vectors and Embeddings

Mikolov, Chen, et al. (2013), Mikolov, Sutskever, et al. (2013), Mikolov, Yih, et al. (2013) suggested an innovative way to create word representations that can be processed in neural network architectures, where the word representations are vectors that encode distributional semantic properties, representing in fact predictive models of contexts for target words.

In general, working with lexical meaning in form of string encodings that are associated with potentially complex feature structures, or using the traditional notation for the meaning of a word like *dog* as *dog'*, is not very useful in computing environments and machine learning. Words and their meaning need to be converted to computable representations, ideally encoded in form of vectors.

The Word2Vec approach proposed in Mikolov, Sutskever, et al. (2013) uses a feed-forward neural network architecture to train vectors for words such that they maximize the prediction of other words in their contexts. I will simplify here somewhat without going into the technical or mathematical details. In common models a vector of a dimensionality of 300 real values is chosen to represent each target word.[2] The vector values are chosen such that the dot-product of such two word vectors represents the likelihood that these two words would occur in a local context of n words or within the set of BoWs for the target word in a large corpus.

The word-vectors are trained using a neural network architecture based on the distributional properties of words in some training corpus. The resulting model is a set of vectors for each word where similar words predict similar words in their context, thus, in terms of geometric similarity or closeness metrics, similar word vectors will be very close to each other.

Computing these word vectors from large corpora might require specific computational resources and memory, and it can be quite time-consuming. Various models are available, pre-computed by the colleagues at Google or Facebook, for example. I describe in the following, how I use such a set of pre-computed word vectors to compute lexical similarity between languages including semantic fields of individual lexical items.

2. This choice is often made empirically, by identifying the vector length that makes the maximizes the accuracy and utility of the model in for example real NLP tasks.

3 Word Vectors of Slavic Languages

The FastText (https://fasttext.cc) provides pre-trained vectors for 157 languages, see Grave et al. (2018), Joulin et al. (2016) for more details. Among those language models are Croatian, Bosnian, Serbian in Latin and Cyrillic alphabet, Serbo-Croatian, Macedonian, and Slovenian vectors. Each vector model consists of a list of word and vector pairs in raw text format. A sample entry would look as follows:[3]

```
godine -0.0186 -0.0258 0.0100 ... -0.0026 0.0066 -0.0221
```

The labeled vectors have a dimension of 300 real numbers. The models, thus, provide us with two sets of information:

1. A list of words per language
2. Vectors for every word that allow us to compute the predicted context words

This enables us initially to compare the lexical inventory of the languages, that is in particular Croatian, Bosnian, Serbian, and Serbo-Croatian.[4]

The nature of the word embeddings allows us to extend the study to a comparison of semantic properties using the predictions of context words for every single lexical item. We expect that there are many lexical overlaps for the Neo-Shtokavian variants spoken in Bosnia, Croatia, and Serbia. In addition to this lexical overlap, we can now compare the semantic fields of the shared vocabulary to study semantic variation and to compute a similarity on the basis of real language use.

The resource is, of course, limited in many ways. The main text-source for computing the word similarities in FastText has been Wikipedia. FastText does not provide any detailed overview of the amount of textual data that has been used for each of the language models. Thus, the size of the lexical inventory for each language, that is the number of types, might vary significantly.

3. The dots representing omitted scalars for space reasons.
4. We assume that the texts that served as the source for the Serbo-Croatian language model have been labeled as such in Wikipedia. The script in the Serbo-Croatian models is mainly using the Latin alphabet.

Another issue is the variation in pronunciation between the variants that is reflected in the orthography. Although the three main languages are variants of Neo-Shtokavian, there are systematic differences in the realizations of certain vowels, reflected in the orthography. For example, in Croatian as the *ije*-kavian variant *-ije-* is realized as in *bijelo* (white), and in Serbian as the *e*-kavian variant as *-e-*, as in *belo* (white). To be able to compare the lexical inventories one needs to utilize some form of automatic orthographic conversion from the *e*-kavian to the *ije*-kavian variant of Neo-Shtokavian, and vice versa. This is also necessary for a study of the context words as representing semantic fields in the sense of Distributional Semantics.

I used automatic orthographic normalization for the lexical inventory comparison. This is, of course, not an ideal approach, since errors can be introduced and some form of false matches could be created. My hope is that applying the conversion bidirectionally, from *e-* to *ije*-kavian and vice versa, the error would be minimized or neutralized.[5]

The orthography-related normalization included *i*-omission in future tense in Croatian, but not necessarily in Bosnian or Serbian, as in *uradit ću – uraditi ću* (I will do it). Certain cases of oppositions were covered as well, as for *-u*/*e* in *porculan* vs. *porcelan*, *-t*/*ć* in *plaća* vs. *plata*. Cases of initial *h*-drop were taken into account, as in *hrđa* vs. *rđa* or *čahura* vs. *čaura*. Also final-*r*-drop was covered, as in *jučer* vs. *juče* or *večer* vs. *veče*. Otherwise the more common phenomena are related to the alternation *je*/*e*, as in *vjetar* vs. *vetar*, and the *ije*/*e* alternation as in *bijelo* vs. *belo*.

There are certain lexical differences that I would not convert, such as the differences in the following word pairs: *riža – pirinač* (rice), *tvornica – fabrika* (factory), *kruh – hleb* (bread), *špinat – spanač* (spinach), *rajčica – paradajz* (tomato). For some of these terms there seems to be a growing shift with respect to lexical familiarity, as for example for *riža – pirinač* (rice). We expect that there is a much larger shift when it comes to the lexical semantics level or constructions with such dissimilar words.

In addition to my own orthographic conversion utilities, I used Cyr-Translit (https://github.com/opendatakosovo/cyrillic-transliteration) by

5. In this study I did not estimate the error introduced by the automatic conversion of the orthography.

the Open Data Kosovo group, which is a free Python module for automatic transliteration.

In addition to the direct vocabulary size comparison, the goal is to study the bags of words that are predicted by the same lexical form across the different languages. Using the FastText model, we can compute the *n* most likely words in the context for the Croatian word *cvijet* (flower) and the Serbian counterpart *cvet*. Each of the words will make predictions of other words in its context. We want to measure the overlap of words among the *n* most likely predicted words in the context BoW.

The hypothesis is that language use differences and differences in the semantic fields of the same concept in two different speaker groups will be reflected in the overlap of words in the context BoW. Since the word embedding models allow us to compute probabilistic distributions of words in the context for any given target word, we could also utilize Information Theoretic or Entropy-based measures to compute a similarity score over the entire distribution, using, for example, Kullback–Leibler divergence or Relative Entropy.

Independent of the final metric, the context BoW items need to be orthographically normalized as well. As mentioned, this likely introduces new margins for error and comes with other sets of issues. Nevertheless, we hope that the tendencies would be significant and clear in the resulting comparison.

4 Results and Discussion

Using the FastText vector models, I compare the lexical overlap and the overlap in the predicted context BoW for each word that looks the same given orthography and orthographic normalization.

From the language models we can derive the number of types in the vocabulary for the different languages. Table 1.3 gives a vocabulary size overview: In the FastText Wikipedia models we observe the vocabulary overlap as represented in Table 1.4, which only shows the real overlap between the language pairs.[6] Table 1.4 shows how both Croatian and

6. A missing overview can be easily generated, which would display the overlap over

language	# types
Bosnian	166,505
Croatian	451,637
Serbian	452,282
Serbo-Croatian	454,674
Slovenian	281,823
Macedonian	176,947

Table 1.3: Number of types in FastText models by language

	Croatian	Serbian	S.-C.	Slovenian	Mac.
Bosnian	137,409	128,417	144,723	71,627	19,584
Croatian		182,784	246,163	107,705	22,885
Serbian			248,644	95,863	22,946
S.-C.				115,293	29,312
Slovenian					23,758

Table 1.4: FastText vocabulary overlap between languages

Serbian are very close to the language model that was labeled Serbo-Croatian. This is in fact more significant than the overlap between Serbian and Croatian as such. There is a slight tendency of the Bosnian model to be closer to Croatian and closest to Serbo-Croatian in terms of lexical overlap.

An interesting analysis is to see how many words from one language can be found in another, and vice versa. Table 1.5 presents these results. Table 1.5 should be read in the following way: for example, in *BS* (Bosnian) we find 30% of words that also occur in the Croatian model of (*HR*).

A discussion of all the predicted word overlaps based on the FastText models goes beyond the scope of this article. I will restrict myself to a brief summary with some examples.

Consider the predicted context words given the word *ban* (governor), as shown in Table 1.6. The example in Table 1.6 provides the best pre-

more than two language groups. I will leave this analysis out for a future publication.

	in BS	in HR	in SR	in SC	in SI	in MK
of BS		82%	77%	86%	43%	11%
of HR	30%		40%	54%	23%	5%
of SR	28%	40%		54%	21%	5%
of SC	31%	54%	54%		25%	6%
of SI	25%	38%	34%	40%		8%
of MK	11%	12%	12%	16%	13%	

Table 1.5: Proportional lexical overlap by language pair

HR	jelačić 0.63265
SR	kuban 0.661549, balaban 0.658383, šaban 0.654212 no jelačić
BS	–

Table 1.6: Predicted context words using FastText

dicted context words and the corresponding probability from the language model.

Providing a complete analysis of the results would be beyond the scope of this article. I hope that the example is convincing that the method to use Word2Vec type of word embeddings for the study of lexical inventories and lexical semantic field variations across languages and dialects can provide interesting and valuable results.

The Word2Vec model and the general approach to compare the sets of predicted context words to compare lexical differences can in fact be expanded and combined with machine translation approaches that normalize the context BoW results. As with the orthographic normalization to align the variation between two languages, we could in fact translate the context BoW words and compare the semantic fields in terms of Distributional Semantics.

As an unsupervised model based purely on quantitative distributional properties of lexical items, this approach has benefits and some serious drawbacks. The problem with the approach to estimate the semantic field of lexical items using the context BoW prediction based on Word2-

Vec models is that it is based on single words only, ignoring idioms and other kinds of multi-word expressions. We do not know to what extent these single word limitations influence the resulting statistics in our case.

The Word2Vec model also ignores lexical ambiguities in that all forms of a word are conflated in one vector. The different meanings and properties of a word like *luka* are a.) port, Nominative Singular or Accusative Plural, b.) onion, Genitive or Partitive, and so on. An approach where different vectors are trained for the different meanings of individual lexical items, potentially using the lemmatized word form, would be more appropriate.

Bibliography

Baroni, Marco & Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4). 673–721.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. In Association for the Advancement of Artificial Intelligence (ed.), *Proceedings of the 14th National Conference on Artificial Intelligence and 9th Conference on Innovative Applications of Artificial Intelligence*, 598–603. Providence, RI: AAAI Press.

Clark, Stephen. 2014. Vector space models of lexical meaning. In Shalom Lappin & Chris Fox (eds.), *The handbook of contemporary semantic theory* (Blackwell Handbooks in Linguistics), 493–522. Hoboken, NJ: Wiley-Blackwell.

Erk, Katrin. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass* 6(10). 635–653.

Firth, John R. 1968. A synopsis of linguistic theory 1930–1955. In F. R. Palmer (ed.), *Selected papers of J. R. Firth 1952–1959*, 168–205. London, England: Longmans.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin & Tomas Mikolov. 2018. Learning word vectors for 157 languages. In Nicoletta Calzolari et al. (eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*. Miyazaki, Japan: European Languages Resources Association.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou & Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. arXiv preprint.

Mikolov, Tomas, Kai Chen, Greg S. Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint.

Mikolov, Tomas, Iljy Sutskever, Kai Chen, Greg S. Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K.Q. Weinberger (eds.), *Proceedings of the 26th Neural Information Processing Systems Conference (NIPS)*. Lake Tahoe, NV: Neural Information Processing Systems.

Mikolov, Tomas, Wen-Tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III & Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. Atlanta, GA: Association for Computational Linguistics.

Turney, Peter & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1). 141–188.