Mapping Phonology to Semantics: A Computational Model of Cross-Lingual Spoken-Word Recognition

Iuliia Zaitova **Badr M. Abdullah** **Dietrich Klakow**
Department of Language Science and Technology (LST)
Saarland Informatics Campus, Saarland University, Germany
{ izaitova | babdullah | dietrich }@lsv.uni-saarland.de

Abstract

Closely related languages are often mutually intelligible to various degrees. Therefore, speakers of closely related languages are usually capable of (partially) comprehending each other's speech without explicitly learning the target, second language. The cross-linguistic intelligibility among closely related languages is mainly driven by linguistic factors such as lexical similarities. This paper presents a computational model of spoken-word recognition and investigates its ability to recognize word forms from different languages than its native, training language. Our model is based on a recurrent neural network that learns to map a word's phonological sequence onto a semantic representation of the word. Furthermore, we present a case study on the related Slavic languages and demonstrate that the cross-lingual performance of our model not only predicts mutual intelligibility to a large extent but also reflects the genetic classification of the languages in our study.

1 Introduction

Speakers of closely related languages are usually capable of understanding each other's speech to a great degree without having a prior exposure to the second language (L2) or switching a lingua franca for communication[1] (Jan and Zeevaert, 2007; Gooskens, 2019). The ability of the listener to comprehend spoken utterances in a different language (L2) using their native language (L1) competence is termed in the sociolinguistics literature as *intercomprehension*. Gooskens (2017) categorized the factors that facilitate intercomprehension into linguistic factors (e.g., inherent cross-linguistic similarity between L1/L2) as well as extra-linguistic factors (e.g., listener's attitude towards communicating in a different language than their own L1).

[1] A language used for communication between people who do not share a native language.

Several studies in the sociolinguistics literature have documented the levels of intercomprehension between related languages through empirical testing of mutual intelligibility with human subjects of different language backgrounds (Gooskens, 2007, 2017; Van Heuven, 2008, *inter alia*). It has been observed that objective measures of cross-language distance—such as lexical distance—are strong predictors of cross-linguistic intelligibility. Therefore, mutual intelligibility of related languages is largely driven by the presence of word cognates—words that encode the same concepts with similar phonological forms across languages.

From the psycholinguistic perspective, the listener's ability to recognize word forms in a different language is an example of the remarkable human ability to cope with the variability of speech (Pisoni and Levi, 2007). Thus, spoken-word recognition across different, but related languages can be considered as lexical access problem—processing the spoken-word form to activate and retrieve the lexical category that is intended by the speaker. In the cognitive modeling literature, the task of spoken-word recognition has been addressed as a mapping problem between an acoustic-phonetic representation of the word form onto its semantic representation in memory (see Weber and Scharenborg (2012) for a detailed overview). Recently, deep neural networks have been explored as models of spoken-word processing and recognition in several studies (Magnuson et al., 2020; Mayn et al., 2021; Matusevych et al., 2021). Our paper adds another contribution to this line of research by considering the cross-lingual aspects of spoken-word recognition and sheds light on its contribution to cross-linguistic intelligibility using a computational model. Our contribution is two-fold: (1) we present a neural model of spoken-word recognition and investigate the degree to which a monolingual model—i.e., has only been trained on a single language—is able to recognize the meaning of spo-

Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 54–63

ken words across related languages, and (2) we present a case study on the Slavic languages which are remarkably similar and mutually intelligible to various degrees. Concretely, we investigate the following research questions:

RQ1 Does the cross-lingual performance of model predict the mutual intelligibility of the languages in our study?

RQ2 Do the results of cross-lingual evaluation reflect the genetic relations among the studied Slavic languages?

RQ3 Which linguistic distance measures predict the cross-lingual performance of the monolingual models? and how do they compare to predictors of human performance?

2 Background and Related Work

2.1 Slavic Intercomprehension

Previous sociolinguistic research on intercomprehension and mutual intelligibility has focused on two related questions: (1) how to experimentally measure the level of mutual intelligibility across related languages using functional testing and human listeners? and (2) which measures of linguistic distance are strong predictors of cross-language intelligibility? (Golubović and Gooskens, 2015). One of the earliest sociolinguistic studies has investigated the intelligibility of Spanish and Brazilian Portuguese (Jensen, 1989). For languages within the Slavic language family, Golubović and Gooskens (2015) have tested mutual intelligibility across two modalities—i.e., text and speech—using three cross-language tasks: (1) word translation, (2) cloze test and (3) picture naming task. Golubović and Gooskens (2015) have observed that the degree of cross-language intelligibility is largely dependent on the genetic proximity of the languages under study. For example, language pairs within the same Slavic sub-family such as Czech and Polish (West Slavic group) are more mutually intelligible than language pairs that cross the group division (Czech and South Slavic languages such as Croatian or Bulgarian). Furthermore, the authors demonstrated that lexical and phonetic similarities across languages are strong predictors of their intelligibility.

Other studies on Slavic intercomprehension take an information-theoretic angle to analyze this phenomenon. For example, Jagrova et al. (2018) investigated the effect of in-context predictability (or lexical surprisal) on the written intelligibility of Czech text for Polish readers and vice versa. Moreover, the information-theoretic metric of word adaptation surprisal has been shown to predict asymmetric intelligibility of Slavic readers of Cyrillic script, namely Russian and Bulgarian (Mosbach et al., 2019). In the speech modality, Kudera et al. (2021) have analyzed the cognate facilitation effect on cross-language auditory lexical processing using a cross-lingual priming study. In summary, the studies we reviewed in this section have demonstrated a great degree of mutual intelligibility among speakers of Slavic languages, and this intelligibility can be predicted by linguistic measures of similarity to a great degree.

2.2 Computational Models of Spoken-word Processing

Using computational models based on deep neural networks (DNNs) to simulate spoken-word processing have been proposed in several prior studies. Magnuson et al. (2020) presented a minimal neural architecture based on an LSTM to map between acoustic word forms onto their respective sparse semantic representation. Mayn et al. (2021) analyzed the effect of speech variability on spoken-word recognition using a DNN model trained on German words from read speech corpora. As part of their experiments, the authors have shown that the model can fairly recognize word cognates from related Germanic languages (namely Dutch and English), and the cross-lingual performance of the model reflected language proximity. Matusevych et al. (2021) introduced a phonetic model of spoken-word processing and demonstrated that the model predicts perceptual difficulties of non-native speakers. It was also shown that neural models of spoken-word processing capture cross-linguistic, typological similarities in their representational geometry (Abdullah et al., 2021b). Macher et al. (2021) proposed a recurrent model that takes as input a phonological sequence and projects it onto a semantic space to investigate orthographic effects on word recognition. These computational studies have demonstrated the usefulness of neural networks to simulate human listeners who have been exposed to a single language, which enables researchers to test specific hypotheses or isolate the effect a particular linguistic level on language processing.

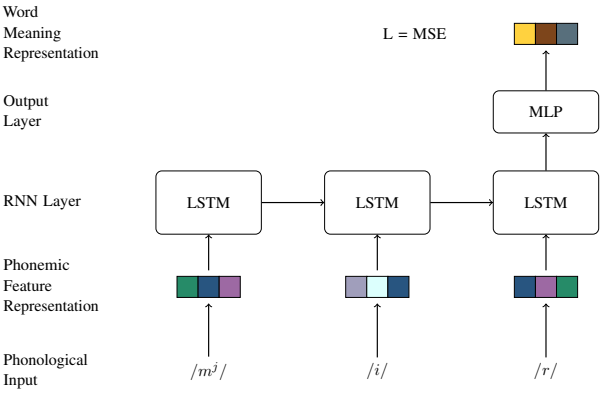

Figure 1: Schematic architecture of the model.

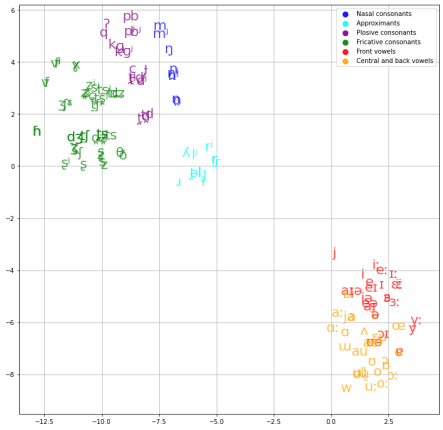

Figure 2: t-SNE visualization of phoneme embeddings vectorized with PHOIBLE feature set. One can notice two clear clusters of consonants (on the left) and vowels (on the right), as well as a visible difference in the positioning of front and back vowels, fricatives, plosives, etc.

3 The Model

Similar to the work of Macher et al. (2021), our model takes a phonological sequence (spoken word form) as input, builds up a whole-word phonological representation of the sequence, and then projects it onto a semantic space (meaning representation) of the lexical item encoded by the word form. Formally, we model the spoken-word recognition task of as a mapping function $\mathcal{F}_{\boldsymbol{\theta}} : \Phi \rightarrow S$, where Φ is the (discrete) space of phonological word forms, $\mathcal{S}$ is the word semantic space, and $\boldsymbol{\theta}$ are the parameters of the mapping function. Since phonological word forms can have any length, we model the function $\mathcal{F}$ using a recurrent neural network (LSTM) followed by a multi-layer perceptron (MLP) (see Figure 1). Given the word form of the lexical category w as a phonological sequence $\Phi(w) = \boldsymbol{\varphi}_{1:\tau} = (\varphi_1, \varphi_2, \ldots, \varphi_\tau)$, a vector representation is computed as

$$\boldsymbol{v} = \mathcal{F}(\boldsymbol{\varphi}_{1:\tau}; \boldsymbol{\theta}) \in R^D \quad (1)$$

Here, D is the dimensionality of the semantic space. Since our goal is to map the phonological input onto a semantic representation, the learning objective is based on vector regression loss and it aims to minimize the term

$$\mathcal{L} = ||\boldsymbol{v} - \Lambda(w)||^2 \quad (2)$$

where $\Lambda(w) \in R^D$ is the ground-truth distributed representation, or semantic word embedding, of the lexical category w. We assume that continuous-space, distributed word representations are available to the model during training.

3.1 Phoneme Representation

Each phoneme in the input phonological sequence $\boldsymbol{\varphi}_{1:\tau} = (\varphi_1, \varphi_2, \ldots, \varphi_\tau)$ is represented as a feature vector based on the PHOIBLE feature set (Moran and McCloy, 2019). That is, we represent each of the 135 phonemes in our inventory as a discrete, multi-valued feature vector of 38 phonetic features similarly to the method introduced in Abdullah et al. (2021a). PHOIBLE includes distinctive feature data for every phoneme in every language. The feature system used is created by the PHOIBLE developers to be descriptively adequate cross-linguistically. In other words, using PHOIBLE feature set allows our model to capture phoneme similarities across languages even if the phonemes have distinct symbols. For each of the 38 available features, every phoneme receives a value, which is $+1$ if the feature is present, -1 if it is not, and 0 if the feature is not applicable.

To illustrate the structure of the phoneme feature representation, we visualize a two-dimensional projection of phoneme representations using the t-SNE algorithm (Van der Maaten and Hinton, 2008) in Figure 2.

3.2 Word Meaning Representation

To represent the word's meaning which our model has to build from the word phonological form, we use distributed word embeddings from fastText (Mikolov et al., 2018). FastText word vectors are pre-trained using the continuous bag-of-words (CBOW) algorithm with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 with contrastive negative sampling.

Figure 3: Major countries where Slavic languages are spoken. Red coloring – for West Slavic, yellow – for Eastern Slavic, and green – for South Slavic.

3.3 Model Hyperparameters and Training

We train six monolingual models for the following languages: Russian, Ukrainian, Polish , Czech, Bulgarian and Croatian. The final model for each language is trained using a batch size of 128 for 150 epochs. We employ the ADAM optimizer (Kingma and Ba, 2014) with the Mean Squared Error (MSE) loss as the vector regression objective function. To account for the different size of input phonemic sequences, we used zero padding to make the size of the input sequence equal to 16. We employ one layer of LSTM, followed by a one-layer MLP consisting of a linear followed by a *tanh* layer. Since every phoneme has 38 features (every phoneme embedding has the length of 38), and every input sequence has the length of 16, the dimensions of the input matrix are 38×16. We use the hidden dimension size of 512, which consequently maps the phonetic sequence to the 300-dimensional target of fastText embeddings. All the models are built using PyTorch (Paszke et al., 2019).

4 Experimental data

In our paper, we present a case study on the Slavic languages which have been shown to exhibit remarkable similarities and high degrees of mutually intelligibility at the conversational level (Sussex and Cubberley, 2006, Golubović and Gooskens, 2015). We use two languages of each of the three main branches of Slavic languages, that is, Russian and Ukrainian for East Slavic; Polish and Czech for West Slavic; and Bulgarian and Croatian for South Slavic[2]. One of the factors that drive our choice is the availability of high quality G2P tools available.

4.1 Phonetic Transcriptions

To obtain an IPA phonetic transcription for each orthographic form of each word embedding in our data, we employ eSpeak speech synthesizer[3]. For the Ukrainian data, we use Epitran transcription library (Mortensen et al., 2018), as this language is not currently supported by eSpeak. For the languages which we only used for evaluation (Belarusian, Slovak, Slovene, Latvian, Romanian, German, and Turkish), the original Northeuralex transcriptions were retrieved using Lexibank (List et al., 2021)[4].

4.2 Training Data

For the training data, we sample experimental word forms from fastText embeddings while excluding the word forms that appear in the test data. Apart from that, we exclude word forms that are classified as parts of speech not present in the test data to reduce noise during training. Parts of speech that are included are *noun, verb, adverb, adjective, pronoun*, and *numeral*.

For each lexical concept in the test data, we make sure that at least three word forms with the same lemma are within the training data. For example, if the word form (ноль, *nolʲ*) is in the test data, it cannot be in the training data, but another word form (ноля, *nolʲa*) can. We hypothesize that the model will be able to capture the semantics of a word by learning to be invariant to inflections and derivations.

4.2.1 Evaluation Data

To evaluate the model performance, we employ parallel lists of word forms from lexicostatistical database NorthEuraLex (Dellert et al., 2019) which cover the 1,016 concepts in all languages. Having a concept for all testing data words in all languages

[2]Henceforth, we use ISO 639-1 codes for the languages: Russian – ru, Ukrainian – uk, Polish – pl, Czech – cs, Bulgarian – bg, Croatian – hr.

[3]http://espeak.sourceforge.net/index.html

[4]Since our input phoneme embeddings capture the features of each phoneme (described in §3.1), transcription difference between the tools should have minimal effect on the model's performance. We additionally tested several transcription tools for the same language, which did not result in a significant change of performance on our model's main task of retrieving meaning of a phonological sequence.

Table 1: Examples of Northeuralex concepts

Concept	Russian Orth	Russian IPA	Czech Orth	Czech IPA	Bulgarian Orth	Bulgarian IPA
EAR	ухо	/u x ɑ/	ucho	/u x o/	ухо	/u x ɔ/
NOSE	нос	/n o s/	nos	/n o s/	нос	/n ɔ s/
FOOD	еда	/je d a/	strava	/s t r a v a/	храна	/x r a n a/
BROTHER	брат	/b r a t/	bratr	/b r a t r/	брат	/b r a t/

allows us to systematically investigate the cross-linguistic performance of the models. Overall, we exclude 514 concepts and use 502 concepts for each of the 13 parallel test sets. Our reasons to exclude some concepts were: 1) the concept does not have a corresponding fastText embedding in any of the 6 training languages; 2) some concepts do not exist in some of the languages as a single word and use a descriptive term for some concepts (for example, the term *breast* corresponds to женская грудь /ʒɛnskəjə gruʲtʲ/) in Russian), which also makes it impossible to retrieve a fastText embedding; 4) a word in one of the 6 training languages maps to more than one concept, which could lead to confusion with its fastText embedding. An example of the NorthEuraLex data we use for testing is represented in Table 1.

5 Evaluation

During testing, the model computes the meaning representation of the phonemic sequence in the test language. To evaluate the model retrieval on the test set, the closest match between the model output and target vector for the model training language is retrieved using cosine similarity. Cosine similarity determines whether two vectors are pointing in roughly the same direction and is measured by the cosine of the angle between two vectors. Cosine similarity, on the abstract level, represents the proximity of the meaning retrieved by the listener to the actual meaning of the word. In other words, it would tell us how semantically similar two given vectors are. Cosine Similarity is computed between a model's output and all the 502 possible ground truth vector representations in the language of training. The vectors to be compared include all the word vectors used for monolingual testing. Given these competing word embeddings, we also calculate average Recall at 1 (R@1), Recall at 5 (R@5), Recall at 10 (R@10), as well as Mean Reciprocal Rank (MRR) for the test data. R@n as the proportion of times that the set of top n word embeddings which are closest to the model's output also includes the ground truth vector representation. If the ground truth is most similar to the output vector of a model, R@1 is 1, otherwise it is 0. Likewise, R@5 is 1, if the corresponding ground truth embedding is within the top 5 most similar words to the output vector, and R@10 is 1 if the embedding is within 10 most similar words. Hence, the average R@n is a number between 0 and 1. The Reciprocal Rank information retrieval measure calculates the reciprocal of the rank at which the first relevant document was retrieved. For evaluation of the test data, we compute an average of Reciprocal Rank for all the given word forms.

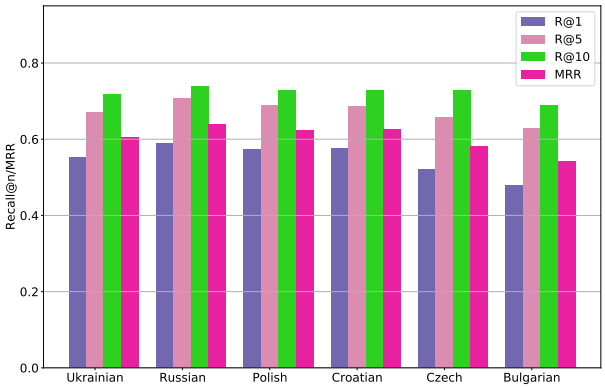

Figure 4: Monolingual performance of the models

5.1 Monolingual Evaluation

The procedures that are used for monolingual and cross-lingual evaluations are comparable, and differ only in the language of the test lexical concept. For both monolingual and cross-lingual evaluations, the retrieved fastTest meaning embeddings for both training and validation sets come from the same embedding space. For the monolingual evaluation, the output embedding for a particular phonemic sequence is compared to groundtruth embeddings of the concepts of test set. The monolingual performance of the models is shown in Figure 4. The monolingual scores for all models are very similar. Such consistency could of course be due to the generally good performance of the current model structure and parameters on any human language. However, it could also be related to structural similarity of the languages of Slavic group (such as, for example, all Slavic languages being synthetic and expressing syntactic relationships via inflection).

5.2 Cross-lingual Evaluation

To make the cross-lingual evaluation comparable across different languages, we compute the cosine

similarity of the L2 target concept to all evaluation concepts in the embedding space of the model training language (L1). For instance, if the model has observed during training the Russian word люди /l^j u d i/ (eng.trans: people), during testing the model on Czech concepts we compute the meaning representation of *lidé* /l i d ɛ/ (eng.trans: people) and then estimate its similarity to test sequences in Russian with the target meaning representation being that of the Russian word люди /l^j u d i/. Such concept mapping during testing has two goals: (1) the pre-trained fastText embeddings for different languages live in different embedding spaces, so it is not possible to compare them as they are, and (2) we assume that a human listener also compares foreign words that they hear to words from their native language, and attempts to retrieve the meaning based on their L1 mental lexicon. For cross-lingual performance, we evaluated each model on all languages under analysis and added three more languages of the Slavic group (East Slavic – Belarusian, West Slavic – Slovak, South Slavic – Slovene) three other languages from the Indo-European language family, to which the Slavic language also belong (German, Romanian, and Latvian), and the Turkish language coming from the Turkic language family[5]. If the model produces human-like behaviour, we can expect it to be better at recognising spoken word forms from more related languages.

The recall at 10 (R@10) results for each model are shown in Figure 5. On the plots, scores for languages of the same language group as the model language, are located on the left side. We also use different color coding for different language group, i.e. reddish colors for East Slavic languages, blueish colors for West Slavic languages, and greenish for South Slavic. Languages outside the Slavic language family are colored in the shades of grey. First, we observe a clear distinction between the retrieval performance of the Slavic and non-Slavic test word forms. The retrieval performance on non-Slavic test word forms (Latvian, Romanian, German, and Turkish) is generally lower for all models except for Bulgarian, which recognizes Romanian evaluation set better than Ukrainian. However, given the geographic proximity between the speaker communities of Romanian and Bulgarian and the fact that both are within the Balkan Sprachbund, this could indicate an effect of lexical borrowing between the two languages. From these findings, we conclude that our hypothesis that the languages which are more genetically related are also more mutually intelligible within the proposed model is mostly supported, with notable exceptions that could related to geographic transfer.

Regarding the evaluation within the Slavic language family, the phonemic sequences in the language from the same subgroup of Slavic languages (such as, Ukrainian for Russian and Croatian for Bulgarian) are recognised significantly better than others by most models. However, there are a few exceptions to this trend. One notable exception in the cross-lingual evaluation is the performance of the Czech model, which seems to have an expected high retrieval performance on Slovak word forms, but unexpectedly does not seem to recognize Polish word forms with a comparable performance. Another surprising result is the fact that the Russian model seems to recognize Croatian and Bulgarian word forms better than Belarusian word forms.

To get further insights onto the cross-lingual performance of the model, we apply hierarchical clustering on the R@10 results between the six models we trained in this study using the Ward algorithm implemented in the SciPy Python library. The Ward's linkage function specifying the distance between two clusters is computed as the increase in the error sum of squares after merging two clusters into a single cluster. The dendrogram of the Ward clustering of R@10 results is shown in Figure 6. The dendrogram in Figure 6 shows we can correctly reconstruct the Slavic language tree from the cross-lingual retrieval performance of the six languages that we have trained models for.

5.3 Correlation with Linguistic Metrics

To investigate which data-driven, linguistic predictors make the model behave as it does, we use Pearson correlation between the cross-lingual model performance and two measures of phonetic-lexical distance. The first metric of phonetic-lexical distance is Levenshtein Distance (LD) where the difference between two strings is calculated as the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. For the second metric, we use Phonologically Weighted Levenshtein Distance (PWLD), which is a measure of phonological similarity between different phonemic sequences or word forms (Fontan et al., 2016). The PWLD met-

[5]the ISO 639-1 codes for the languages: Belarusian – be, Slovak – sk, Slovene – sl, German – de, Romanian – ro, Latvian – lv, Turkish – tr.

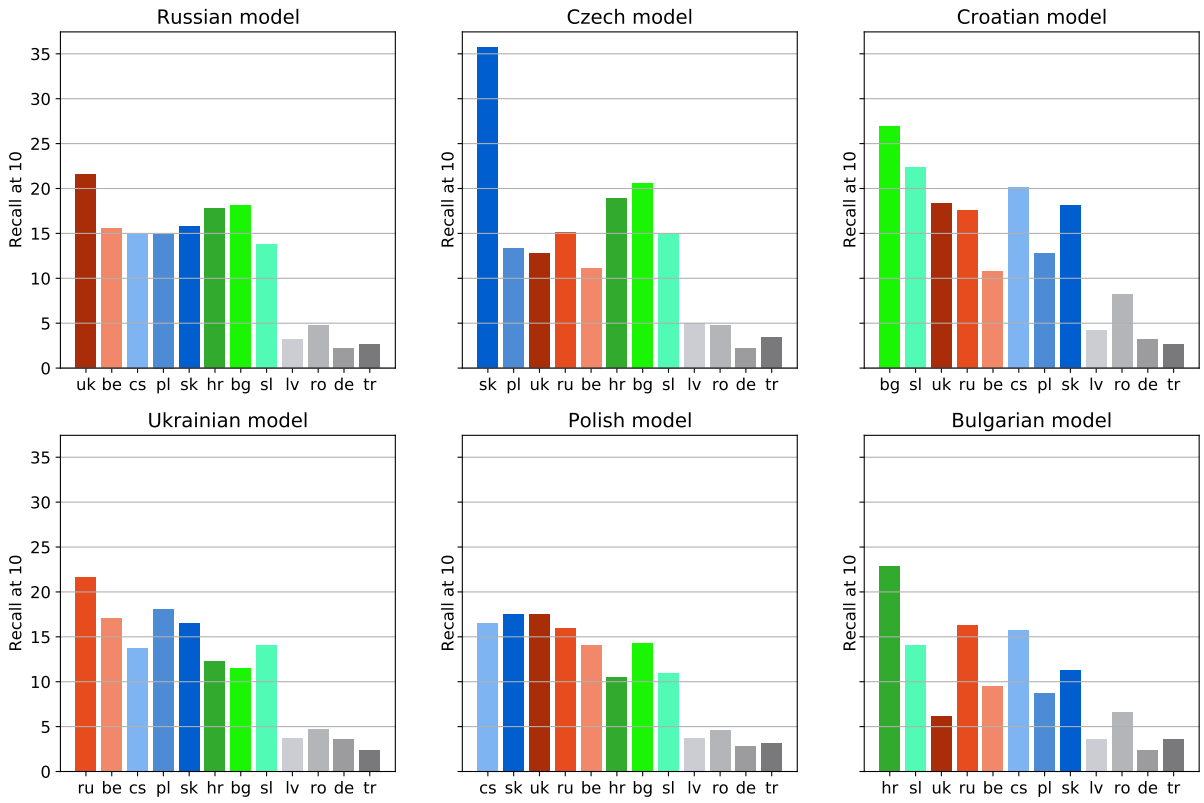

Figure 5: Recall at 10 results. Each plot corresponds to a model trained on one language, while y-axis shows evaluation languages. ISO 639-1 codes for the languages: Ukrainian – uk, Russian – ru, Belarusian – be, Czech – cs, Polish – pl, Slovak – sk, Croatian – hr, Bulgarian – bg, Slovene – sl, Latvian – lv, Romanian – ro, German – de, Turkish – tr.

Figure 6: Dendrogram of the Ward clustering of R@10 results.

ric is an extension of the string-based Levenshtein distance that also calculates the cost of each phone substitution based on phoneme features. We suppose that PWLD is more suitable for cross-lingual analysis than Levenshtein Distance, since it is more capable of catching less apparent phonological similarities, such as, for example in the pair of Czech and Bulgarian cognates *ucho /u x o/* and ухо */u x ɔ/*, where phonemes */o/* and */ɔ/* are very similar to each other. We use the same adaption of the original PWDL proposed in Abdullah et al. (2021a) to make it suitable for our analysis.

Table 2: Pearson correlation coefficient for metrics under analysis. Statistical significance is marked with * and *** for $p < 0.05$ and $p < 0.001$, respectively.

	R@10	MRR	cos sim	LD	PWLD
R10		0.98***	0.5***	-0.74***	-0.57***
MRR			0.5***	-0.75***	-0.56***
cos sim				-0.29*	-0.44***
LD					0.8***
PWLD					

Table 2 shows the correlation scores of all the metrics under analysis. We observe that both metrics correlate with MRR and R@10, while the correlation with cosine Similarity scores are much lower. Surprisingly, PWLD has a lower correlation with the retrieval metrics than LD, even though it uses the same phoneme vectorization scheme as the model.

5.4 Qualitative Analysis

Figure 7 shows t-SNE visualization on the output on the Russian model. For t-SNE computation, we used output vectors for all the test data. On

the visualization, only the concepts *FOG*, *WIND*, *FISH*, and *MOSQUITO* are shown. For concepts *WIND*, *FISH*, and *MOSQUITO* one can observe clear clusters of concepts, as they also appear to sound similarly in all the 6 languages. This is not the case with the concept *FOG*. As shown in Figure 7, t-SNE clustered the concept in different languages quite far from each other even for similarly sounding words. It is interesting that concepts *MOSQUITO* and *WIND* that do not sound similar, but probably have a contextual, distributional similarity, appear close to each other. This probably has to do with the nature of the target fastText embeddings, which are trained to predict the word's context. Additionally, we provide the top retrieved words for the model trained on Russian and tested on Ukrainian. Table 3 demonstrates other candidates in Ukrainian for some phonemic sequences in Russian. The English translation of the concept is given in the brackets.

From the lists of cross-lingual nearest neighbors reported in Table 3, one can notice that the model learns to push semantically similar words closer to each other, despite them having a very different phonetic shape (for instance, *soup-porridge-food* or *who-why-was*). This could again be related to be the nature of fastText embeddings (Mikolov et al., 2018) that we used as target embeddings for the model. As already mentioned, the vector for each word also contains information about this word's context. As a result, the output embeddings produced by the model for contextually close words appear to have a lot in common and are recognized as semantically similar.

Another observation from Table 3 is the clear advantage of shorter and non-content spoken word forms over longer ones. Most of the short words in the list are non-content words, that do not have any distinctive semantic context, and appear in any type of text. In this regard, these words can be seen as items that share fewer features compared to longer words and content words.

6 Discussion and Conclusion

In this paper, we presented a spoken-word recognition model based on a recurrent neural architecture that maps variable-length phonological sequences of word forms into their respective meaning representations. Our goal is to simulate auditory lexical processing in human listeners where we test the model on word forms from closely related

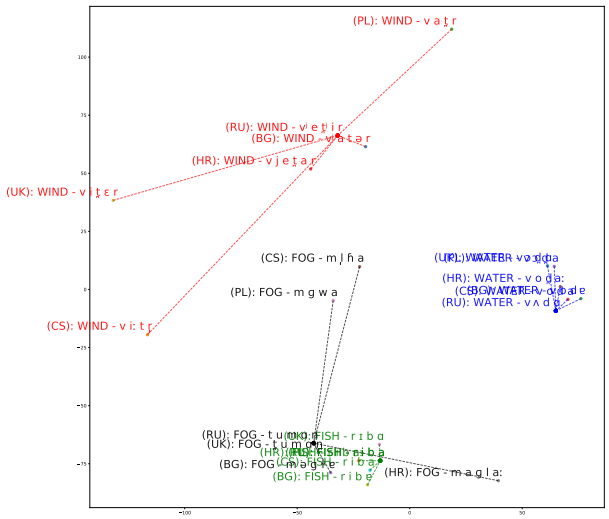

Figure 7: t-SNE on the concept retrieval of the Russian model.

languages and investigate the cross-lingual performance of the model. Furthermore, we presented a case study on the family of Slavic languages, which are known to be remarkable similar and exhibit (partial) mutual intelligibility to various degrees. We grounded our research on the findings from the sociolinguistics literature of Slavic mutual intelligibility and intercomprehension. Using our proposed model, we trained different instances of our model on six Slavic languages: Bulgarian, Croatian, Czech, Polish, Russian, and Ukrainian. Finally, we conducted a cross-lingual evaluation on our trained models to investigate their performance on retrieving and recognizing word forms from other L2 languages.

Returning to our research questions in §1, the cross-lingual analysis of our model performance has shown a trend where the model performance is better on languages that exhibit higher cross-linguistic intelligibility as documented in sociolinguistics studies (**RQ1**). However, this effect is more consistent within South and East Slavic languages, but less consistent in the case of West Slavic languages (Czech and Polish). The factors that drive this inconsistency remain unknown and would require further future work to identify and analyze. Despite this inconsistency, the clustering analysis on the cross-lingual concept retrieval performance resulted in a dendrogram that reflects the traditional genetic classification of the six studied Slavic languages onto West, East, and South languages (**RQ2**). Furthermore, we have shown that cross-linguistic phonetic-lexical similarities

Table 3: Top scored candidates in Ukrainian for the model trained on Russian

	/j a/ (‘I’)	**/r ɑ n a/ (‘wound’)**	**/k t o/ (‘who’)**	**/k ɑ ʃ a/ (‘porridge’)**	**/s u p/ (‘soup’)**
Nearest neighbors	/j A/ (‘I’)	/r ɑ n ɑ/ (‘wound’)	/x t / (‘who’)	k ɑ ʃ ɑ (‘porridge’)	s u p/ (‘soup’)
	/d ɛˈ/ (‘yes’)	/j ɑ/ (‘I’)	/t u t/ (‘here’)	/r ɑ n ɑ/ (‘wound’)	/k ʃ / (‘porridge’)
	/s i m/ (‘if’)	/ ɑ p k ɑ/ (‘hat’)	/v r / (‘whisper’)	/v ɔ r ɔ / (‘whisper’)	/d ɛˈ n/ (‘day)
	/x t ɔ/ (‘who’)	/j i ɑ/ (‘life’)	/t ɔ m u/ (‘why’)	/ʃ ɑ p k ɑ/ (‘hat’)	/x r t/ (‘food’)
	/j i ɑ/ (‘life’)	/d ɛˈ/ (‘yes’)	/b i j/ (‘was’)	/k n ɑ/ (‘book’)	/k r^{j}uk/(‘$hook$’)

between the languages—operationalized as string and feature-based phonetic distance on a parallel word list—correlate with the cross-lingual concept retrieval performance of the model. This finding is consistent with the observation in the sociolinguistics literature regarding how lexical similarity between languages facilitates intercomprehension (e.g., the cognate facilitation effect). Therefore, the cross-lingual concept retrieval performance of our model can be predicted using measures of linguistic distance similar to those that predict cross-language comprehension performance (**RQ3**).

The work presented in this paper can be further extended in different directions. For instance, mutual intelligibility between related languages have been found in many cases to be asymmetric. For example, speakers of Portuguese seem to understand Spanish better than the other way around. Future work could analyze and investigate whether or not and to what extent such an asymmetric behavior is observed in our model.

7 Acknowledgements

The authors would like to thank the anonymous reviewers for their encouraging feedback and insightful comments on the paper. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

References

Badr M. Abdullah, Marius Mosbach, Iuliia Zaitova, Bernd Möbius, and Dietrich Klakow. 2021a. Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study. In *Proceedings of Interspeech 2021*, pages 4194–4198.

Badr M. Abdullah, Iuliia Zaitova, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. 2021b. How familiar does that sound? cross-lingual representational similarity analysis of acoustic word embeddings. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 407–419, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Johannes Dellert, Thora Daneyko, and Alla et al. Münch. 2019. Northeuralex: a wide-coverage lexical database of northern eurasia.

Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, and Xavier Aumont. 2016. Using phonologically weighted levenshtein distances for the prediction of microscopic intelligibility. In *Annual conference Interspeech (INTERSPEECH 2016)*, page 650.

Jelena Golubović and Charlotte Gooskens. 2015. Mutual intelligibility between West and South Slavic languages. *Russian Linguistics*, 39:351–373.

Charlotte Gooskens. 2007. The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and multicultural development*, 28(6):445–467.

Charlotte Gooskens. 2017. Dialect intelligibility. *The handbook of dialectology*, pages 204–218.

Charlotte Gooskens. 2019. Receptive multilingualism. *Multidisciplinary perspectives on multilingualism*, pages 149–174.

Klara Jagrova, Tania Avgustinova, Irina Stenger, and Andrea Fischer. 2018. Language models, surprisal and fantasy in Slavic intercomprehension. *Computer Speech Language*, 53.

D Jan and Ludger Zeevaert. 2007. *Receptive multilingualism: Linguistic analyses, language policies and didactic concepts*, volume 6. John Benjamins Publishing.

John B. Jensen. 1989. On the mutual intelligibility of Spanish and Portuguese. *Hispania*, 72(4):848–852.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

Jacek Kudera, Philip Georgis, Bernd Möbius, Tania Avgustinova, and Dietrich Klakow. 2021. Phonetic distance and surprisal in multilingual priming: Evidence from Slavic. In *Interspeech*, pages 3944–3948.

Johann-Mattis List, Robert Forkel, Simon J Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D Gray. 2021. Lexibank: A public repository of standardized wordlists with computed phonological and lexical features.

Nicole Macher, Badr M. Abdullah, Harm Brouwer, and Dietrich Klakow. 2021. Do we read what we hear? modeling orthographic influences on spoken word recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 16–22, Online. Association for Computational Linguistics.

James S Magnuson, Heejo You, Sahil Luthra, Monica Li, Hosung Nam, Monty Escabi, Kevin Brown, Paul D Allopenna, Rachel M Theodore, Nicholas Monto, et al. 2020. Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive science*, 44(4):e12823.

Yevgen Matusevych, Herman Kamper, Thomas Schatz, Naomi H Feldman, and Sharon Goldwater. 2021. A phonetic model of non-native spoken word processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Alexandra Mayn, Badr M. Abdullah, and Dietrich Klakow. 2021. Familiar words but strange voices: Modelling the influence of speech variability on word recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 96–102, Online. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Marius Mosbach, Irina Stenger, Tania Avgustinova, and Dietrich Klakow. 2019. incom.py - a toolbox for calculating linguistic distances and asymmetries between related languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 810–818, Varna, Bulgaria. INCOMA Ltd.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

David B Pisoni and Susannah V Levi. 2007. Some observations on representations and representational specificity in speech perception and spoken word recognition. *The Oxford handbook of psycholinguistics*, pages 3–18.

Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge University Press.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vincent J Van Heuven. 2008. Making sense of strange sounds:(mutual) intelligibility of related language varieties. a review. *International journal of humanities and arts computing*, 2(1-2):39–62.

Andrea Weber and Odette Scharenborg. 2012. Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3):387–401.