UNIVERSITÄT
DES
SAARLANDES

Master's Thesis

Modeling Cross-Language Spoken Word Recognition with Neural Networks: a Case Study on Slavic Languages

submitted by

Iuliia Zaitova

First Supervisor: Prof. Dr. Dietrich Klakow

Second Supervisor: Badr M. Abdullah

Department of Language Science and Technology

23 May 2022

Declaration of Authorship

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Ich versichere, dass die gedruckte und die elektronische Version der Masterarbeit inhaltlich übereinstimmen.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged. I assure that the electronic version is identical in content to the printed version of the Master's thesis.

In Saarbrücken date 23.05.2022 ..... Author's signature

Acknowledgements

First and foremost, I am very grateful to my supervisors Badr M. Abdullah and Prof. Dr. Dietrich Klakow for their their assistance, guidance, and patience throughout this whole work.

I also want to thank all my friends and family that provided me with support and encouragement during the time of writing my thesis.

Abstract

Speakers of closely related languages are usually capable of (partially) comprehending each other's speech without explicitly learning the other language. Empirical testing of such mutual intelligibility, as documented by several studies in the sociolinguistics literature, has been shown to be non-trivial. That is, the speakers' contact with the foreign language biased the results, giving rise to asymmetrical intelligibility, where speakers of one language can understand speakers of a related language better than vice versa. In this thesis, we propose an alternative methodology to overcome this bias using a computational model to simulate cross-language lexical processing. Our proposed spoken-word recognition model is based on a multi-layer Long Short-Term Memory (LSTM) neural network that maps variable-length phonemic representations of wordforms into their meaning representations. We show that a model which has only been exposed to one of the six Slavic languages (Bulgarian, Croatian, Czech, Polish, Russian, and Ukrainian) is able to partially recognize spoken wordforms in the other five languages under analysis. To a large extent, our experimental results reflect the linguistic proximity of Slavic languages and compare to the results of previous sociolinguistic studies.

Contents

1 Introduction

Oftentimes, speakers of closely related languages are capable of understanding some information from each other's speech without actually learning the other language or using any lingua franca. C. Gooskens et al. (2018) call such a communication strategy receptive multilingualism [11], [37], [5]; it is documented across various language families throughout time and in various social settings [37]. Previous sociolinguistic research has focused on the following questions in relation to this phenomenon: How exactly do we measure the level of mutual intelligibility cross-linguistically [11], [10]? How can we compare receptive multilingualism among different languages [11], [10]? Psycholinguistic research regards receptive multilingualism as the problem of lexical access, where wordforms serve as access codes with which a listener can access semantic forms in their memory [12], [34], [38], [34]. In the field of computational linguistics, the most relevant question is whether a computational model would be able to simulate the effect of receptive multilingualism, which is also a question that the current thesis aims to deal with [20], [36].

Computational modeling of spoken word recognition has been previously investigated by several authors. Macher et al. (2021) [18] explore the influence of orthography on speech recognition. They train neural network models on spoken word forms in German to test whether they can map an unseen lemma to its correct meaning representation. Mayn et al. (2021) [20] model human speech perception using word-aligned read speech data in German and test, as part of their experiments, whether the model which has only been exposed to German, is able to recognize cognates in two related languages, English and Dutch. In this thesis, we also present a cross-linguistic study, but we do not only focus on phonetically similar spoken wordforms like cognates or different lemmas of the test items. We use more diverse training data with wordforms that resemble the spoken language exposure of a native speaker, and consider six different languages. We only look at the languages of Slavic group which are considered to be closely related and moderately mutually intelligible at the conversational level [35].

The language group under consideration is Slavic languages. In particular, we are looking at six Slavic languages from three different sub-groups of Slavic languages: Czech and Polish for West Slavic, Russian and Ukrainian for East Slavic, and Bulgarian and Croatian for South Slavic. Slavic languages are considered to be moderately mutually intelligible, at least at the conversational level [35]. This fact, together with the previous sociolinguistic studies on cross-language intelligibility between Slavic languages, makes us more confident about choosing this language group to effectively simulate cross-lingual spoken word recognition.

The primary goal of this thesis is to computationally test whether and to what extent a computational model which has only been exposed to one language would be able to recognize the meaning of spoken words in closely related languages. We try to approach

this goal considering not only computational, but also sociolinguistic and psycholinguistic aspects of it.
This thesis seeks to explore the following research questions in the relation to the primary goal:

- Are our computational modeling results comparable to the previous sociolinguistic studies on intelligibility?
- Do the results reflect the geneological classification of languages? What role does the relatedness of languages play in the ability of a monolingually trained model to recognize words in other languages?
- Since we mostly aim to simulate the effect, and not the process of human language learning, what features of the model and the input cause the model's behaviour? For instance, would the model be sensitive to the average word length of the input?

The rest of this thesis work explores the answers to these questions and is organized as follows: §2 introduces relevant previous research and explains important concepts employed in the work; §3 gives an account of the methodology, describes the model's architecture, as well as training and testing procedures; §4 outlines both quantitative and qualitative experiments we have conducted, and §5 provides a discussion of the obtained results. Finally, we summarize the results and restate the main points, indicating opportunities for future research in §6.

2 Related Work

In this section we introduce important concepts that can facilitate the understanding of this thesis work, and describe previous research in several areas that are relevant or closely related to Cross-Language Spoken Word Recognition. Although the core of the thesis lies in simulating word recognition with neural networks (NN), we attempted to approach this Machine Learning task from a linguistic perspective and put a strong consideration to previous sociolinguistic and psycholinguistic studies on the subject of Cross-Language Intelligibility.

Experiments on the topic of Spoken Word Recognition that are relevant to this work can be broadly divided into linguistic experiments employing real human subjects, and computational simulations of word recognition using Machine Learning models. In the first subsection of this chapter, we focus on relevant sociolinguistic notions and experiments on mutual intelligibility. In the second subsection, we describe the process of lexical access from psycholinguistic perspective and introduce several studies attempting to simulate it computationally. Finally, in the third subsection, several NN-based computational models of speech recognition in connectionist studies are presented.

2.1 Sociolinguistic Studies of Mutual Intelligibility

In the first chapter, when talking about cross-language intelligibility, we introduced the term receptive multilingualism and described it as the phenomenon that occurs when speakers of different languages understand each other without knowing the foreign language.

Jan D. ten Thije, and Ludger Zeevaert, the editors of the book 'Receptive multilingualism' (2007), define the phenomenon as the language ordinance in which interlocutors use their respective mother tongue when speaking to each other [37]. Giving a broader definition of the term, Gooskens, C. et al. (2018) highlight that receptive multilingualism is based on the fact that some language pairs are so closely related that the speakers are able to communicate with each other using their mother tongue without prior instruction in the foreign language [11]. Focusing purely on language similarity gives a rise to a question — 'What makes two different languages mutually intelligible?' and brings us back to another question asked previously in the introduction — 'How exactly do we measure the level of mutual intelligibility cross-linguistically?'. These are the two concerns we need to address before modeling cross-linguistical word recognition.

Mutual intelligibility, methods of measuring and predicting it received most attention in previous sociolinguistic studies. According to Golubović, J. and Gooskens, C. (2015) [10], what determines how much a speaker of language A can understand the speaker of language B includes the percentage of cognates (words with a common origin which often have a similar form) between the languages, the similarity of their phonological, or-

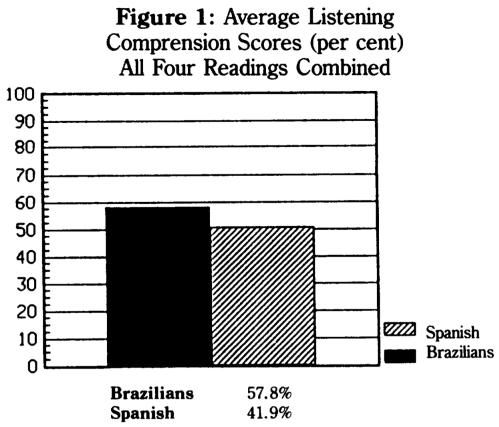


Figure 1: Results of J. B. Jensen (1989)'s experiments on listening-comprehension task

thographic, morphological and syntactic systems, as well as some extra-linguistic factors (for instance, attitude to the language and amount of contact with it). J. Golubović differentiates between inherent intelligibility, which is based purely on the similarities between the related languages in question (for instance Slovak speakers reading or listening to Bulgarian for the first time in their lives) and acquired intelligibility, which refers to intelligibility of a language acquired over time through formal education or other forms of exposure. Acquired intelligibility without any inherent intelligibility occurs in the case of completely unrelated languages, such as Russian and Estonian.

One of the works pursuing to establish the level of mutual intelligibility is J. B. Jensen [14]. The experiments described in it measure the intelligibility of Spanish and Brazilian Portuguese. Two listening-comprehension audio recordings (one in Spanish and one in Portuguese), each with the length of around three minutes, are used. The participants are asked to first listen to an audio recording in the foreign language (Portuguese for Spanish speakers and Spanish for Portuguese speakers), and then answer five written multiple-choice comprehension questions in their native language. Additionally, participants fill in a short questionnaire about their attitude towards the foreign language and their exposure to it. The results of the experiment confirmed that the two languages are mutually intelligible to an extent of around 50% to 60%. However, as shown in Figure 1, the Brazilian group achieved a higher average score than the Spanish group (58% versus 50%, the difference is significant on both the Pearson's correlation test and the two-tailed t-test for means). Although such results support the common belief that Portuguese speakers understand Spanish better than vice-versa, the correlation between the experiment results appear to have a high correlation (.304 at the .01 level of significance) with the previous contact of the participants with the foreign language (according to the additional questionaire). That may suggest that speakers of Portuguese perform better at understanding Spanish not because of their inherently better comprehension ability, but rather due to a high exposure to Spanish newscasts and television (as is proposed in the paper).

Another study by Golubović, J. and Gooskens, C. (2015) [10] uses three different methods to empirically test the intelligibility between languages of two Slavic language branches, West (Czech, Slovak, and Polish) and South (Croatian, Slovene, and Bulgarian). The paper only focuses on these two branches because it covers the Slavic languages spoken in the European Union, which does not include Russian, Belarusian, and Ukrainian. The authors test intelligibility in both written and spoken language using three experimental tasks — word translation task, cloze test, and picture task. They hypothesize that the distinction between two branches would be kept in the results, for instance, that a speaker of Slovak will always understand Polish and Czech better than any South Slavic language.

An illustration of the result of the spoken word translation task are shown in Figure 2, where the language distance data is plotted in two-dimensional space. In this version of the task, participants listen to 50 words (each word repeated twice), randomly chosen from a 100 word list. They are given 10 seconds to translate each word. One can notice that the results correspond the expectations set by the authors: from the plot, the division between West and South is quite obvious.

Despite such generally good results, there was an extralinguistic issue in the experiments that became apparent through the results of cloze task. In this task, some words in a text are deleted and replaced by gaps. The participants have to insert the correct words into the gaps. In Figure 3, we can see an asymmetry in intelligibility of some languages, especially in Croatian-Slovene pair, where Slovene speakers can understand written and spoken Croatian better than vice versa ($t = -6.561$, $p < 0.001$). Just like the results of the previously described experiments of J. B. Jensen [14], it might potentially be related to the fact that speakers of Slovene are more likely to be exposed to Croatian in their lifetime than vice versa.

Although one can see that previous exposure to the language plays a certain role in mutual language intelligibility, it is not clear how exactly and to what extent language exposure and other extralinguistic factors influence the understanding of the foreign language. Consequently, we also do not know how the differences in language structure affect intelligibility and have a difficulty estimating what J. Golubovic calls 'inherent intelligibility', based purely on the similarities of the related languages [9].

To address this issue, Gooskens, C. et al. (2018) [11], that investigate the mutual intelligibility between languages of Europe, propose eliminating extralinguistic factors by employing tabula rasa speakers that have never been exposed to each other's language before. The first subjects of this kind that come to mind could be children that never heard or learned foreign languages before. Unfortunately, these speakers are much harder to find and an experiment with them would not be the most straightforward one to organize.

As the second solution suggested by the authors, one can simulate an unexposed language user using a computational model [11], which would be the focus of the rest of this thesis.

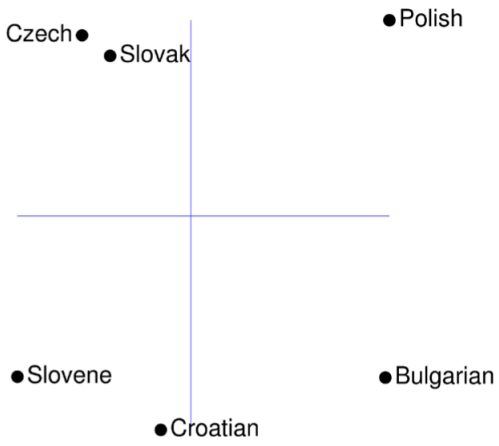


Figure 2: Results of Golubović, J. and Gooskens, C. (2015)'s experiments on spoken word translation task

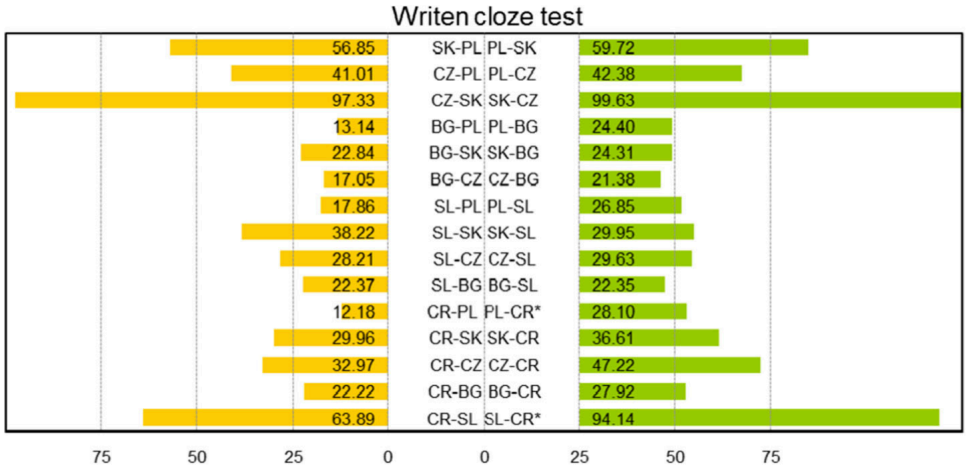


Figure 3: Results of Golubović, J. and Gooskens, C. (2015)'s experiments on written cloze task

2.2 Lexical Access

When trying to model spoken word recognition, we need to have a clear idea of what exactly we are trying model and what is the process that we are trying to simulate. Psycholinguistic research generally agrees that in order to comprehend a word, listeners access semantic forms in their memory [12], [34], [38].

According to Strijkers, K. and Costa, A. (2011), phonemes extracted from an acoustic signal, are access codes to words that are mapped to semantic representations in our mental lexicon (our pool of mentally stored information) [34]. Hence, when doing computational modeling, we need to simulate the process of accessing lexical knowledge as output given an acoustic realization of a word form as input.

The idea of modeling this process computationally is not new and has been especially

	Test language					
	Croatian	Slovene	Bulgarian	Czech	Slovak	Polish
Participants' native language						
Croatian		63.89	22.22	32.97	29.96	12.18
Slovene	94.14		22.37	28.21	38.22	17.86
Bulgarian	27.92	22.35		17.05	22.84	13.14
Czech	47.22	29.63	21.38		97.33	41.01
Slovak	36.61	29.95	24.31	99.63		56.85
Polish	28.10	26.85	24.40	42.38	59.72	

Figure 4: The results of the written cloze test (Golubović, J. and Gooskens, C., 2015)

popular in connectionist literature.

2.3 Distributed Word Representations

As the first step in simulating the process of lexical access, one needs to model some initial word representation (word shape). One of the works that attempted it, is Pinter, Y. et al. (2017) [29], that presents an approach to infer out-of-vocabulary word embeddings from pre-trained, limited-vocabulary models mapping from spellings to distributed word embeddings from Polyglot [3].

Using word embeddings helps to generalise over lexical features by placing each word in a lower-dimensional space. Distributed word embeddings, such as embeddings from Polyglot, that the paper uses, are real valued vector representations that capture semantic and syntactic features. The authors assume there is generative wordform-based process for creating these word embeddings. They train a model over the existing vocabulary with pre-trained embeddings using a Bidirectional Long Short-Term Memory (LSTM) architecture model shown in Figure 3, and use the trained model for predicting the embedding of unseen words. The results of this implementation are quite peculiar. In Figure 6, you can see selected English out-of-vocabulary words with their nearest in-vocabulary words computed by cosine similarity.

From these examples the researchers make several conclusions:

- the model learns word shape well (acronyms, capitalizations)
- part-of-speech is learned across multiple suffixes (pesky – euphoric, ghastly)
- word compounding is detected (e.g., lawnmower – bookmaker, postman)
- semantics are not learned well (as is to be expected from the lack of context in training, and probably, no inflections of the same lemmas seen by the model during training).

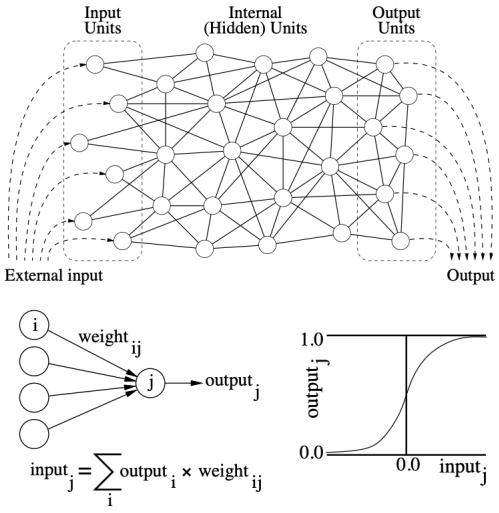


Figure 5: A generic connectionist model from Plaut, D. C. (2000) [30] Input units receive input and send connections to internal units that, in turn, send connections to output units. The activity of each unit is a nonlinear function of the summed weighted input from other units. The resulting activity over the output units constitutes the network's output.

2.4 Computational Modeling of Spoken-Word Recognition

To model a listener of a language, as is required in the case of our work, accessing lexical knowledge given a specific acoustic needs to be modeled.

Relevant recent work in this area is focused on mapping a phonological representation of a word form to its semantic embedding. A lot of this work comes from connectionist studies that focus on learning based upon graded, malleable, distributed representations. Connectionist models take cognitive processes in the form of cooperative and competitive interactions among large numbers of simple, neuron-like processing units (see Figure 5). Learning in such models involves modifying the values of connection weights based on feedback from the environment on the accuracy of the system's responses [30].

A major characteristic of the connectionist approach to language is that it applies the very same processing mechanisms that are used across the full range of linguistic structure [32]. Though units and connections in connectionist models are not generally considered to be in one-to-one correspondence with actual neurons and synapses, these systems attempt to capture the essential computational properties of the vast ensembles of real neuronal elements found in the brain, through simulations of smaller networks of units [30]. An example of a notable connectionist model is that of Gaskell and Marslen-Wilson, 1997 [8], which represents the process of spoken word recognition as a mapping of low-level acoustic features onto the stored semantic and phonological representations. Other influential spoken-word recognition models include, among others, the Cohort model [19], and the Shortlist model [26].

A big inspiration for our model is the work of Macher, N. et al. (2021) [18], which is also a connectionist example of a spoken word recognition model. The paper introduces Long

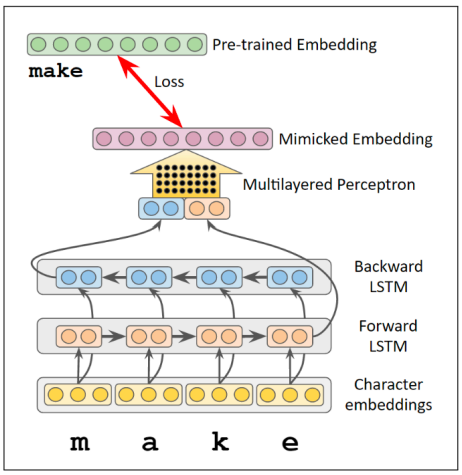


Figure 6: Pinter, Y. et al (2017)'s model for predicting the embedding of an unseen word [29]

OOV word	Nearest neighbors	OOV word	Nearest neighbors
MCT	AWS OTA APT PDM SMP	compartmentalize	formalize rationalize discern prioritize validate
McNeally	Howlett Gaughan McCallum Blaney	pesky	euphoric disagreeable horrid ghastly horrifying
Vercellotti	Martinelli Marini Sabatini Antonelli	lawnmower	tradesman bookmaker postman hairdresser
Secretive	Routine Niche Turnaround Themed	developiong	compromising inflating shrinking straining
corssing	slicing swaying pounding grasping	hurtling	splashing pounding swaying slicing rubbing
flatfish	slimy jerky watery glassy wrinkle	expectedly	legitimately profoundly strangely energetically

Figure 7: Results of Pinter, Y. et al (2017)'s model for predicting the embedding of an unseen word [29]

Short-Term Memory (LSTM) architecture models of spoken word recognition exploring the influence of orthography on speech recognition. In contrast to previously discussed Pinter, Y. et al. (2017) [29], the authors of this paper train their models to map not from spellings of words, but from phonetic sequences, which simulates the spoken-word recognition process, similarly to what we want to achieve in this thesis. They implement two models of spoken-word recognition, one of which is schematized in Figure 7. The purpose of experiments in this work is to make the models learn the meaning of spoken words seen during training by mapping phonetic sequences to distributed word embeddings and generalize to similar but unseen words.

For example, by training the German language model on inflected forms, Maus (mouse), Mäuse (mice) and Häuser (houses), one can afterward test whether the model can get to the correct meaning representation of an unseen lemma like Haus (house). Both models described in the paper achieve significant results and perform well in word meaning retrieval with the best model (the online model) achieving Recall@10 = 0.7, where Recall@10 is defined as the proportion of times that the top 10 set of word forms whose word embeddings are closest to the output of the model also includes the actual word that the model has processed.

A distinctive feature of our work is that we try to simulate spoken word recognition cross-

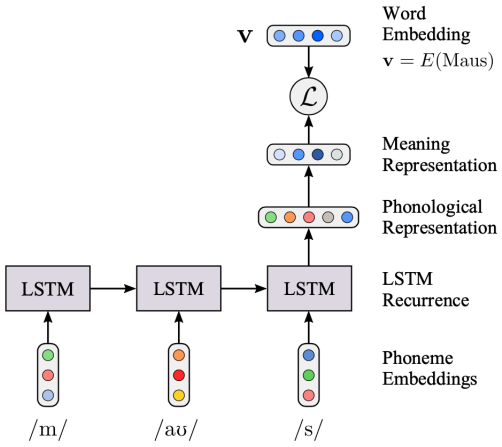


Figure 8: Macher, N. et al. (2021)'s model of spoken word recognition. First, the model takes the respective phonological word form as input. Then, it should build a vector representation that corresponds to a phoneme sequence, to then produce a word meaning representation as output. This meaning representation should be as close as possible to the actual ground truth embedding of the phonological word form [18].

lingually. However, all of the previous connectionist works discussed so far present monolingual models — those that are trained to recognize out-of-vocabulary words or spoken word forms on the same language. A notable work that also incorporates cross-lingual testing is Mayn, A. et al. (2021) [20]. The paper uses deep neural network with several convolutional and several linear layers (Figure 9) to model human speech processing using word-aligned read speech data in German. The objective of the model is to project the spoken words into the word embedding space in such a way that an embedding and an utterance representing the same word will be encouraged to end up close together. On an abstract level, the authors try to simulate a listener as a mechanism which learns to associate wordform with meaning. As part of their experiments, the researchers test whether the model which has only been exposed to German is able to recognize cognates in two related languages, English and Dutch. For that, they pass English and Dutch cognates of seen German words through the model and retrieve the top n closest word embeddings in the shared semantic space. The resulted multilingual listener model is evaluted by its embedding retrieval scores. In the result, the model appears to be relatively good at recognizing cognates in the two related languages, with R@10 = 20% for Dutch and R@10 = 11% for English. According to the paper, the performance reflects genealogical relatedness between the three languages, and could be thought of as reflecting intelligibility of related Germanic languages based purely on linguistic similarity factors. The research conducted in our work tries to simulate cross-language intelligibility employing six Slavic languages in a similar fashion by direct mapping of acoustic wordforms and meaning embeddings.

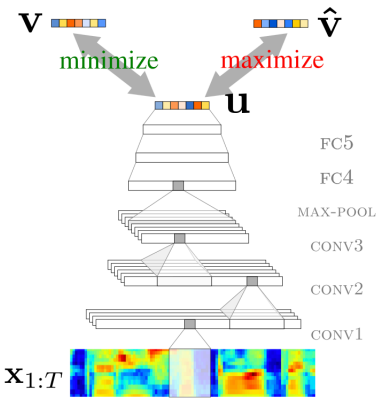


Figure 9: Mayn, A. et al. (2021)'s model of human speech perception [20]. A vector of MFCCs which were extracted from the audio is passed through three convolutional layers with batch normalization and ReLU, followed by max pooling, two fully connected layers and ReLU, and finally a linear projection, which outputs a vector of the same dimensions as the word embeddings.

3 Methodology

Having described all the necessary background knowledge in the Chapter 2, we can proceed with the main goal of this thesis, which is to train and computationally test whether and to what extent a computational model which has only been exposed to one language would be able to recognize the meaning of spoken words in closely related languages.

3.1 Proposed Model

Our proposed spoken word recognition model is based on a four-layer Long Short Term Memory (LSTM) neural network. By mapping a phonetic sequence input onto a meaning representation, it is targeting a vector regression problem.

3.1.1 Architecture and Hyperparameters

After some experiments with the network's structure and hyperparameters, we chose the best performing model, whose architecture is schematically depicted in Figure 10. All the six models for the six corresponding Slavic languages are trained using the same architecture and hyperparameters.

The final model for each language is trained using a batch size of 128 for 150 epochs, since there is hardly any improvement when training the model further. We employ the ADAM optimizer [15], and the Mean Squared Error (MSE) loss as the loss function for training.

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 two-layer feedforward network consisting of one linear and one tanh layer. Each layer of the used model is described in detail below.

In its basic architecture, LSTM, the first part of our model, has three gates: Input Gate, Forget Gate, Output Gate. For each element in the input sequence, each layer computes the following Function 1:

$$
\begin{aligned}
\boldsymbol{i_t} &= \sigma(\boldsymbol{W_{ii}x_t} + \boldsymbol{b_{ii}} + \boldsymbol{W_{hi}h_{t-1}} + \boldsymbol{b_{hi}}) \\
\boldsymbol{f_t} &= (\boldsymbol{W_{if}x_t} + \boldsymbol{b_{if}} + \boldsymbol{W_{hf}h_{t-1}} + \boldsymbol{b_{hf}}) \\
\boldsymbol{g_t} &= tanh(\boldsymbol{W_{ig}x_t} + \boldsymbol{b_{ig}} + \boldsymbol{W_{hg}h_{t-1}} + \boldsymbol{b_{hg}}) \\
\boldsymbol{o_t} &= (\boldsymbol{W_{io}x_t} + \boldsymbol{b_{io}} + \boldsymbol{W_{ho}h_{t-1}} + \boldsymbol{b_{ho}}) \\
\boldsymbol{c_t} &= \boldsymbol{f_t} \circ \boldsymbol{x_t} + \boldsymbol{i_i} \circ \boldsymbol{g_t} \\
\boldsymbol{h_t} &= \boldsymbol{o_t} \circ tanh(\boldsymbol{c_t})
\end{aligned}
\tag{1}
$$

$\boldsymbol{h_t}$	hidden state at time t
$\boldsymbol{c_t}$	cell state at time t
$\boldsymbol{x_t}$	input at time t
$\boldsymbol{h_{t-1}}$	hidden state of the layer at time t-1
$\boldsymbol{i_t, f_t, g_t, o_t}$	input, forget, cell, and output gates
σ	sigmoid function
$\circ$	Hadamard product

As the second component of our model, we employed a multi-layer perceptron (MLP) with tanh activation function, the formula is demonstrated below under Equation 2.

$$\underset{1\times n}{\boldsymbol{v}} = tahn(\underset{1\times m}{\boldsymbol{h_\tau}}\ \underset{m\times n}{\boldsymbol{W_{MLP}}} + \underset{1\times n}{\boldsymbol{b}}) \tag{2}$$

$\boldsymbol{n}$	300
$\boldsymbol{m}$	512
$\boldsymbol{h_\tau W_{MLP} + b}$	linear transformation
$\boldsymbol{h_\tau}$	output of the hidden layer of LSTM
$\boldsymbol{W_{MLP}}$	weight matrix
$\boldsymbol{b}$	bias (vector)

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 38x16. 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 [27].

3.1.2 Phoneme Representation

When it comes to the input of the following models, each phoneme in a phonetic sequence is represented by a phoneme embedding. For vectorizing the input phonemic sequence, we represent each of the 135 phonemes in our inventory as a discrete, multi-valued feature vector based on the PHOIBLE feature set [24], similarly to Abdullah, B. M. et al. (2021) [2].

PHOIBLE dataset 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 similarity across languages even if phonemes differ in their graphemic representation.

For each of the 38 available features, every phoneme receives a value, which is '+' if the feature is present, '-' if it is not, and '0' if the feature is not applied. An example is

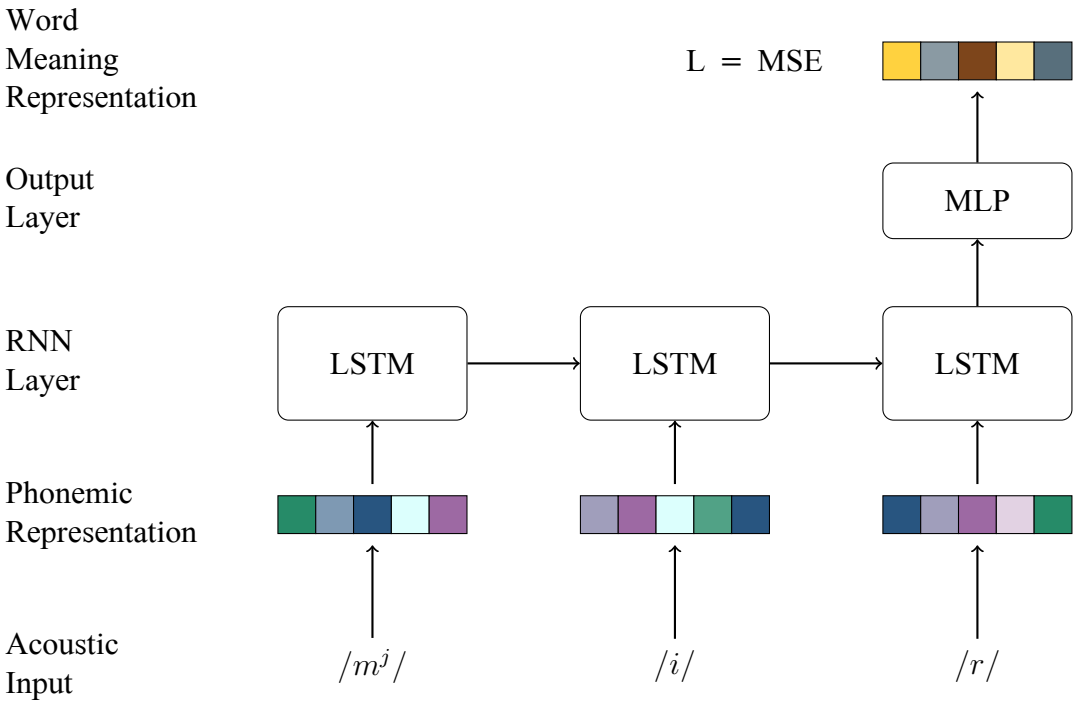


Figure 10: Schematic architecture of the model

Table 1: Example of PHOIBLE phoneme representation (only 7 of 38 features are shown)

Name	tone	stress	syllabic	short	long	consonantal	sonorant	continuant
a	0	-	+	-	-	-	+	+
aː	0	-	+	-	+	-	+	+
j	0	-	-	-	-	-	+	+
k	0	-	-	-	-	+	-	-

shown in the Table 1.

To turn this representation into vectors which are eventually projected into meaning embeddings, we substited '+', '-', and '0' with numerical values of 1, -1, and 0. To better represent the structure of the embeddings, we used t-SNE, a technique to visualize high-dimensional data by giving each datapoint a location in a two or three-dimensional map. T-SNE helps to reveal the structure of the data on different scales. The t-SNE visualization of phoneme embeddings vectorized with PHOIBLE feature set is shown in Figure 11.

3.1.3 Word Meaning Representation

To represent the word's meaning which our model has to learn, we used word embeddings drawn from fastText. FastText is a library by the Facebook AI Research (FAIR) lab for efficient learning of word representations and sentence classification [23]. It combines concepts introduced by the natural language processing and machine learning communities in the last few decades. The word vectors, which are available for download from the official fastText website, were pre-trained using CBOW with position-weights, in

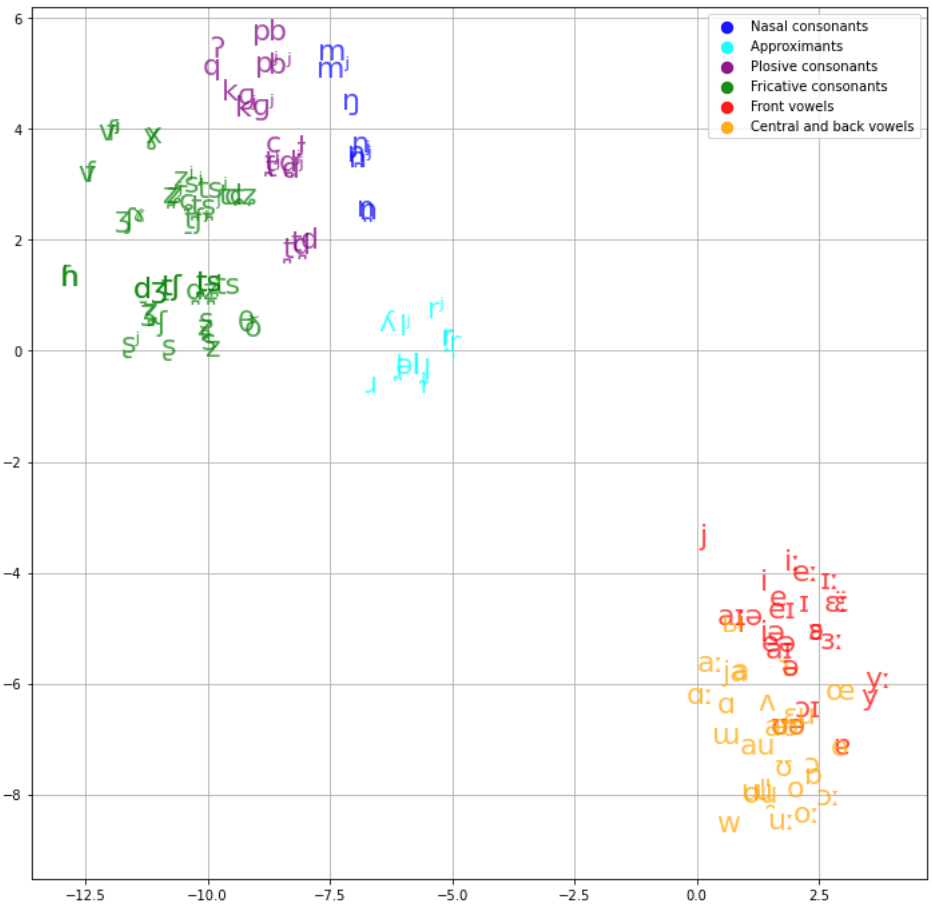


Figure 11: t-SNE visualization of phoneme embeddings vectorized with PHOIBLE feature set. One can notice two clear clusters of consonants (on the left) and vowels (on the right), as well as a visible difference in the positioning of front and back vowels, fricatives, plosives, etc.

dimension 300, with character n-grams of length 5, a window of size 5 with contrastive negative sampling. In the CBOW model, the distributed representations of context (or surrounding words) are combined to predict the word in the middle. In this sense, the embeddings from fastText aim to capture the semantic meaning of words.

For each of the six monolingually trained models, we used 50,000 word embeddings for the training data, around 450 word embeddings for testing data, and around 250 for development data.

3.1.4 Training and Loss Function

Similarly to the previously described model of Macher, A. et al. (2021) [18], our model accepts a phonemic sequence (spoken wordform) as input, builds first a whole-word phonological embedding of the phonetic sequence, and then projects it onto a semantic embedding (meaning representation) of a lexical item represented by this wordform. During training, the network tries to minimize the distance between the computed representation and the fastText embedding of the word using Mean Squared Error (MSE) loss function. MSE is one of the most commonly used loss function for regression. It is calculated as the mean overseen data of the squared differences between true and predicted values, defined as follows:

$$MSE = ||\boldsymbol{E}(\boldsymbol{w}) - \boldsymbol{v}||_2 \tag{3}$$

MSE	mean squared error
$\boldsymbol{E}(\boldsymbol{w})$	estimated vector
$\boldsymbol{v}$	target vector

3.2 Experimental Data

Following the traditional geneological classification of languages, Slavic languages are divided into three branches: West Slavic, South Slavic and East Slavic.

For the languages under analysis, we chose to 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 [1]. Our choice of these exact languages is also driven by the availability of high quality G2P tools available. We aim to create six models corresponding to the six languages we chose to use. We assume that each model trained to understand a particular language (L1) would also be able to partially understand the test data in other languages that are closely related to L1.

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

Figure 12: Major countries where Slavic languages are spoken. Red coloring – for West Slavic, yellow – for Eastern Slavic, and green – for South Slavic

3.2.1 Obtaining Transcriptions

To transliterate orthographic text of our data as IPA, we employ eSpeak speech synthesizer [1]. For the IPA transcription of Ukrainian data, Epitran transcription library [25] is used, as this language is not supported by eSpeak. On its turn, Epitran does not support Bulgarian, so using the same automatic transcription tool turned out to be impossible [2]. For the languages which we only used for testing (Belarusian, Slovak, Slovene, Latvian, Romanian, German, and Turkish), the original Northeuralex transcriptions were retrieved using Lexibank [16].

3.2.2 Training

For the training data, we draw a required number of wordforms that have a FastText embeddings randomly, excluding the wordforms that appear in the test data. Apart from that, we exclude wordforms 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 lexeme in the test data, we add from one to three wordforms with the same lemma into the training data. For example, if the word form (ноль, n o lʲ) is in the test data, it cannot be in the training data, but another word form (ноля, n o lʲa) can. According to our assumption, this can help the model to deduce the semantics of a word by learning to ignore inflections and affixes.

3.2.3 Evaluation

For testing, we employ parallel lists of wordforms from lexicostatistical database NorthEuraLex [6] which cover the same 1,016 concepts in all languages. Having a concept for all testing data words in all languages allows us to compare these word lists to each other and assess the mutual intelligibility of the models.

- One issue that we faced when retrieving the data is that some concepts do not exist in some of the languages, so they had to be excluded from the testing data for all languages.
- Some languages use a descriptive term instead of one word for some particular concepts (for example, the term breast corresponds to женская грудь /ʒɛˈnskəjə grutʲ/) in Russian), which we also decided to avoid.
- The variations from 450 in the testing data size did not exceed 10% for any of the training language. An example of the NorthEuraLex data we use for testing is represented in Table 3.2.3.

[2] To make sure that using different transcription tools does not skew the performance of our models, we tested several transcription tools for the same language, which did not result in the change of performance on our model's main task of retrieving meaning of a phonetic sequence.

Concept	Russian Orth-c form	IPA	Czech Orth-c form	IPA	Bulgarian Orth-c form	IPA
EAR	ухо	/u x a/	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/

3.3 Evaluation Procedure

During testing we compute 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 found 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, as defined below:

$$\cos(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\langle \boldsymbol{y}, \hat{\boldsymbol{y}} \rangle}{\|\boldsymbol{y}\| \|\hat{\boldsymbol{y}}\|} = \frac{\sum_{i=1}^{n} \boldsymbol{y_i} \hat{\boldsymbol{y}}_i}{\sqrt{\sum_{i=1}^{n} (\boldsymbol{y_i})^2} \sqrt{\sum_{i=1}^{n} (\hat{\boldsymbol{y}}_i)^2}} \tag{4}$$

cos	Cosine Similarity
$\|\boldsymbol{y}\|$	Euclidean norm of observed vector
$\|\hat{\boldsymbol{y}}\|$	Euclidean norm of estimated vector
$\langle \boldsymbol{y}, \hat{\boldsymbol{y}} \rangle$	Inner product of observed and estimated vector

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 450 possible ground truth vector representations in the language of training (around 450 vectors). 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. Similarly, 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 wordforms.

3.3.1 Monolingual Evaluation

The procedures that are used for monolingual and cross-lingual evaluations are slightly different. For monolingual evaluation, the fastTest meaning embeddings for both training and validation sets come from the same embedding space. Thus, the output embedding for a particular phonemic sequence is compared to groundtruth embeddings from the same language test set, among which the embedding for the input phonemic sequence is also present. The results of monolingual evaluation are demonstrated in the Subsection 4.

3.3.2 Cross-lingual Evaluation

Comparison among the output embeddings for cross-lingual evaluation is less trivial, because in this case the training data and validation data come from different embedding spaces (fastTest vectors are trained independently for each language). However, since we have a parallel list of concepts, we use cosine similarity to compare between the output and the target representation of the sequence with the same concept in the training language. For instance, if we trained the model on Russian word люди /l' u d i/ (eng.trans: people), and test it on Czech, we would compute the meaning representation of lidé /l i d ə/ (eng.trans: people) and then we would test its similarity to test sequences in Russian with the target meaning representation being that of the Russian word люди /l' u d i/. Such concept mapping during testing has two goals: firstly, the pre-trained FastText embeddings for different languages come from different embedding spaces, so it is not possible to compare them as they are; secondly, we assume that a human listener also compares foreign words that they hear to words from their native language, and tries to understand the meaning based on their already existing mental lexicon. The results of cross-lingual evaluation are demonstrated in the Chapter 4.2.

4 Testing and Experiments

The following section elaborates on the relevant experiments conducted to understand what and how the trained model learns.

We begin with the experiments on model structure which allowed us to choose the best performing model. In these experiments, we explore different values for such hyperparameters and techniques as:

1. Size of training data
2. Batch size
3. Type of LSTM architecture
4. Type of phoneme embeddings used as input to the model
5. Dropout regularization

We proceed with multilingual experiments and analysis, designed to assess the model's performance in response to input data from languages of different degree of relatedness. Finally, we look more closely at the relationship between the input data and output generated by the model, using both quantitative analysis and detailed qualitative analysis. Such analysis aims to give us a better understanding of how the trained model's performance compares to cross-linguistic spoken word recognition of a human listener.

4.1 Experiments on Model Structure

4.1.1 Training and Batch Size

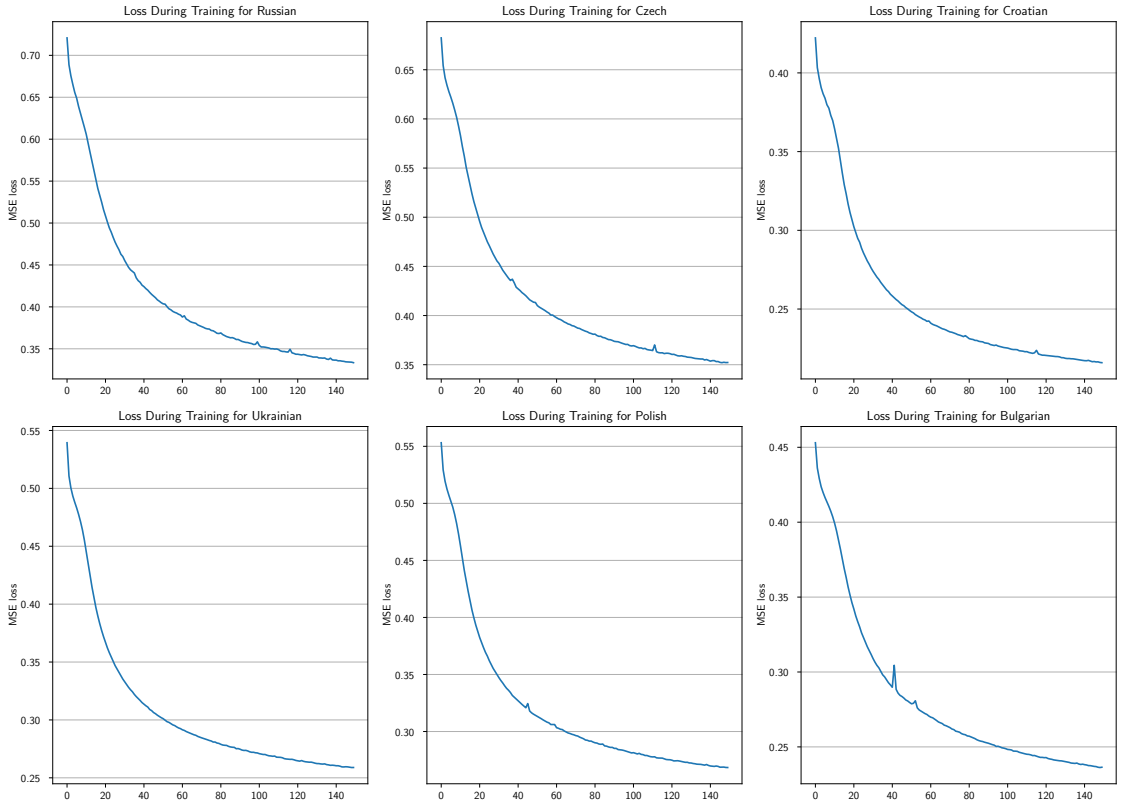
As part of the model development, we assessed whether increasing the training data size has a significant influence on the test results. We trained the models with the training size of 10,000, 20,000, and 50,000 (the largest number that is possible to retrieve from the available resources).

The set of 50,000 word-embedding pairs produced the best results and gave on average 15.7% monolingual performance enhance compared to the the training size of 10,000. The decrease of training loss by epoch for all models is demonstrated in Figure 13.

4.1.2 Unidirectional vs. Bidirectional LSTM

We additionally assessed whether the model's performance on both monolingual and cross-lingual evaluation would increase when using bidirectional instead of unidirectional LSTM. However, since we did not obtain any significant performance increase, we decided to proceed with the unidirectional LSTM model, which is also a more cognitively plausible model of human listening.

Figure 13: Training loss for all languages



4.1.3 Embedding Type

The model's performance is also assessed according to the type of phoneme embeddings used. The embedding types that we experimented with are:

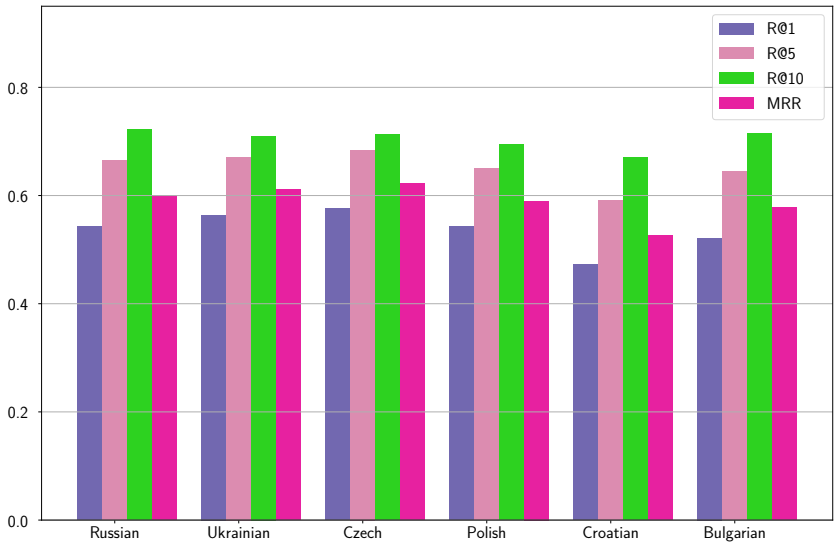
1. Randomly initialized embeddings created with PyTorch Embedding class (without using pretrained embeddings)
2. Embeddings created by vectorizing phoneme features PHOIBLE's phonological inventory data (described in detail in 3.1.2)
3. Embeddings created by using the Word2Vec algorithm that trains a neural network to reconstruct linguistic contexts (Mikolov et al. (2013) [22]) on phonemes.

The types of embeddings we used with their R@10 scores for cross-lingual testing (model trained on 50,000 spoken Russian wordforms, tested on Ukrainian data) are provided below in Table 2. Embeddings created with PHOIBLE's phonological inventory data produced the best results.

Table 2: R@10 by different Embedding Types

Embedding Type	R@10 for testing of Russian model on Ukrainian data
Randomly initialized	9.5%
PHOIBLE	23.4%
Word2Vec	10.5%

Figure 14: Monolingual performance of the models



4.1.4 Dropout Regularization

To make sure that the high difference between mono- and cross-lingual performance are not due to overfitting, we experimented with adding dropout p of 0.1, 0.2, and 0.5 between LSTM and feedforward network. As dropout only decreased the model's monolingual performance and did not change the cross-lingual performance, we decided to not employ it during training.

4.2 Multilingual Experiments

We have evaluated the monolingually trained models on both monolingual and cross-lingual performance.

The monolingual performance of the models is shown in Figure 14.

For cross-lingual performance, we added three more languages of the Slavic group (East Slavic — Belarussian, West Slavic – Slovak, South Slavic – Slovene) three other languages from the Indo-European language family (German, Romanian, and Latvian), and the Turkish language coming from the Turkic language family [3]. 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 1, 5, and 10 results for each model are given in Figures 15, 16, and 17 correspondingly. Mean Reciprocal Rank results are shown in Figure 18. For comparison, the results on recall at 10 for randomly shuffled input sequences are shown in Figure 19.

[3] the ISO 639-1 codes for the languages: German – de, Romanian – ro, Latvian – lv, Turkish – tr

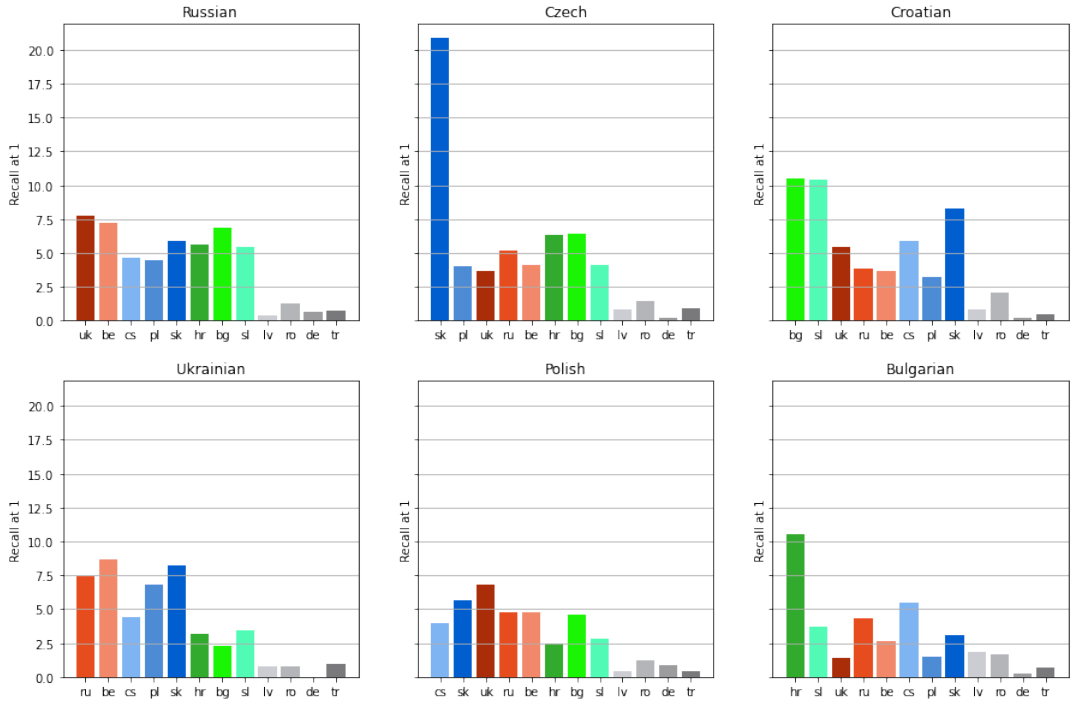


Figure 15: Recall at 1 results. Each plot corresponds to a model trained on one language. Each plot corresponds to a model trained on one language. ISO 639-1 codes for the languages: Ukrainian – uk, Russian – ru, Belarusian – be, Czech – cs, Polish – pl, Slovak – sk, Croatian – hr, Bulgarian – bg, Slovene – sl, Latvian – lv, Romanian – ro, German – de, Turkish – tr.

On the plots, scores for languages of the same language group as the model language, are located on the left side. We also used 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 of Slavic group are colored in the shades of grey.

From the figures, we can see that adding these languages indeed shows us a clear superiority of the performance on the languages from the Slavic group over less related ones, and, in some cases, the superiority of the performance on the same branch of Slavic languages.

As a way of classifying the performance of models trained on different languages, we used R@10 results to apply hierarchical clustering with Ward's algorithm using SciPy Python library v1.8.0. 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 below in Figure 20. To test the statistical significance of the difference in performance on different languages, we conducted the bootstrap test, which is thoroughly described in [4], on the R@10 results. The bootstrap estimates p-value x though a combination of simulation and approximation, drawing many simulated test sets x_i and counting how often an accidental advantage of $sigma_x$ or greater is seen. The estimated p-values for

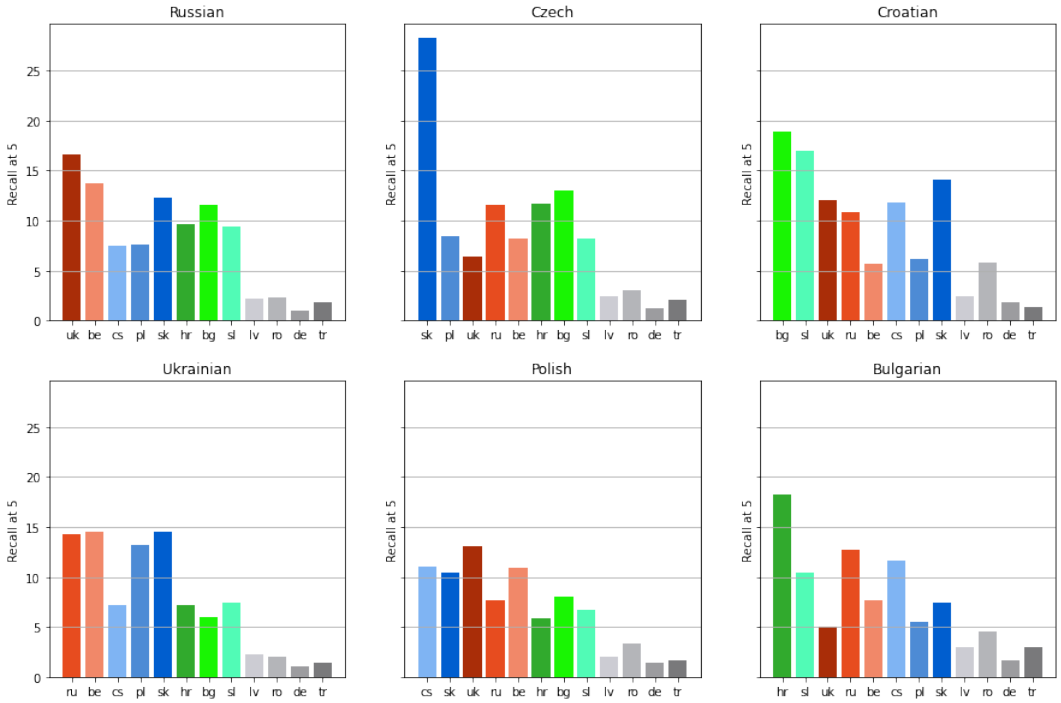


Figure 16: Recall at 5 results. Each plot corresponds to a model trained on one language. ISO 639-1 codes for the languages: Ukrainian – uk, Russian – ru, Belarusian – be, Czech – cs, Polish – pl, Slovak – sk, Croatian – hr, Bulgarian – bg, Slovene – sl, Latvian – lv, Romanian – ro, German – de, Turkish – tr.

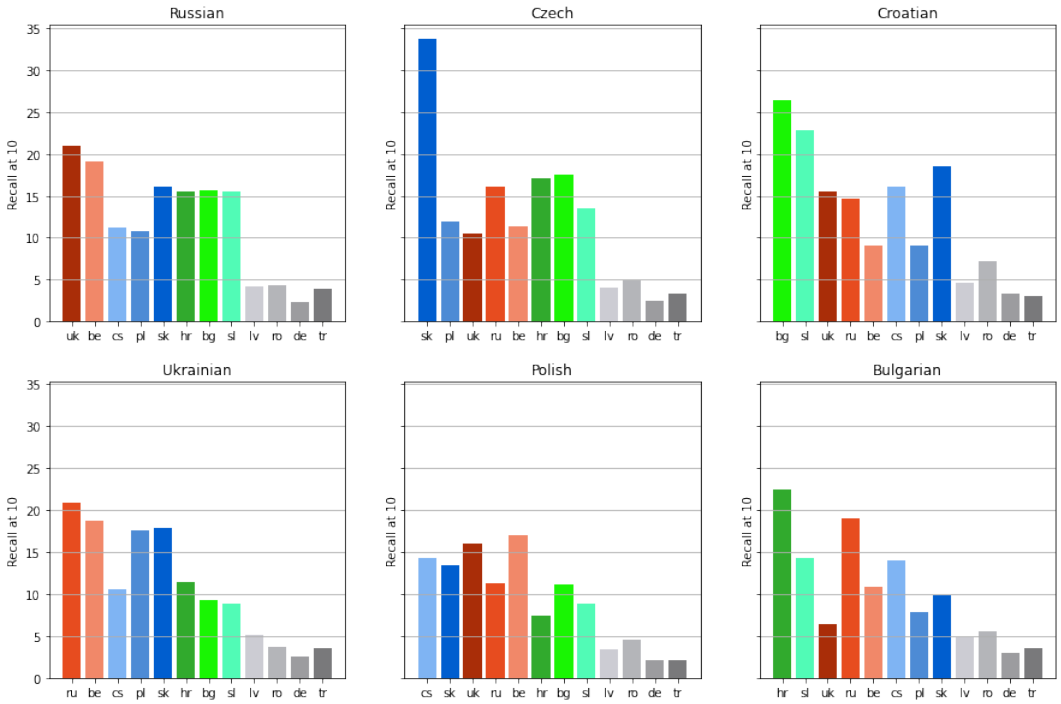


Figure 17: Recall at 10 results. Each plot corresponds to a model trained on one language. ISO 639-1 codes for the languages: Ukrainian – uk, Russian – ru, Belarusian – be, Czech – cs, Polish – pl, Slovak – sk, Croatian – hr, Bulgarian – bg, Slovene – sl, Latvian – lv, Romanian – ro, German – de, Turkish – tr.

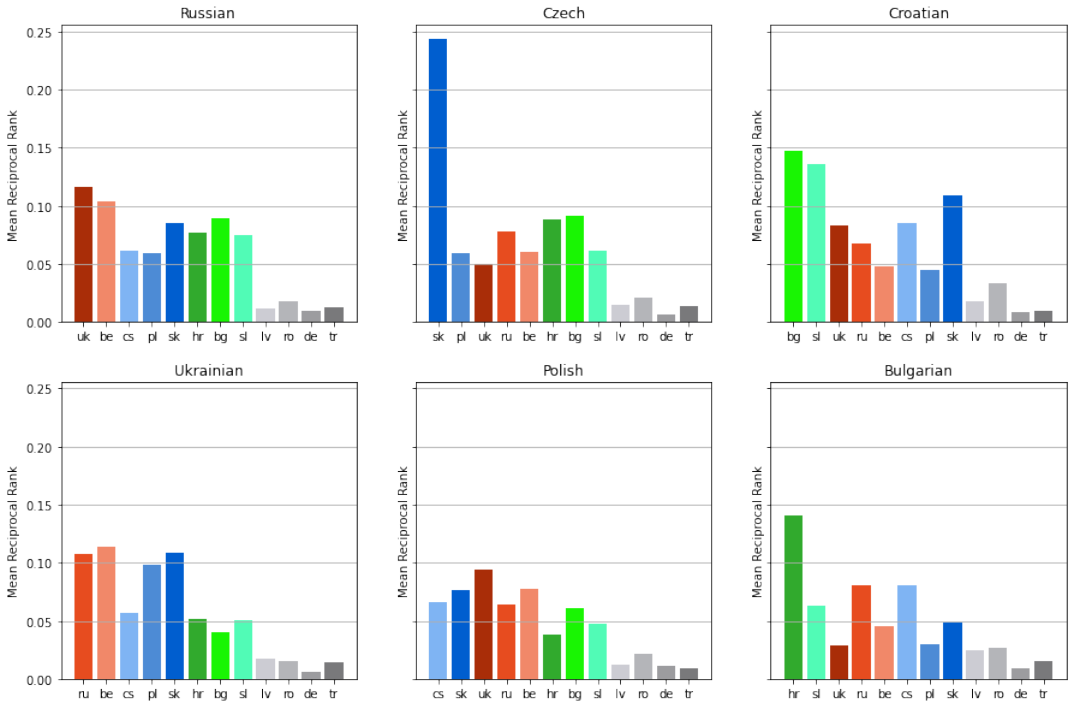


Figure 18: Mean Reciprocal Rank results. Each plot corresponds to a model trained on one language. ISO 639-1 codes for the languages: Ukrainian – uk, Russian – ru, Belarusian – be, Czech – cs, Polish – pl, Slovak – sk, Croatian – hr, Bulgarian – bg, Slovene – sl, Latvian – lv, Romanian – ro, German – de, Turkish – tr.

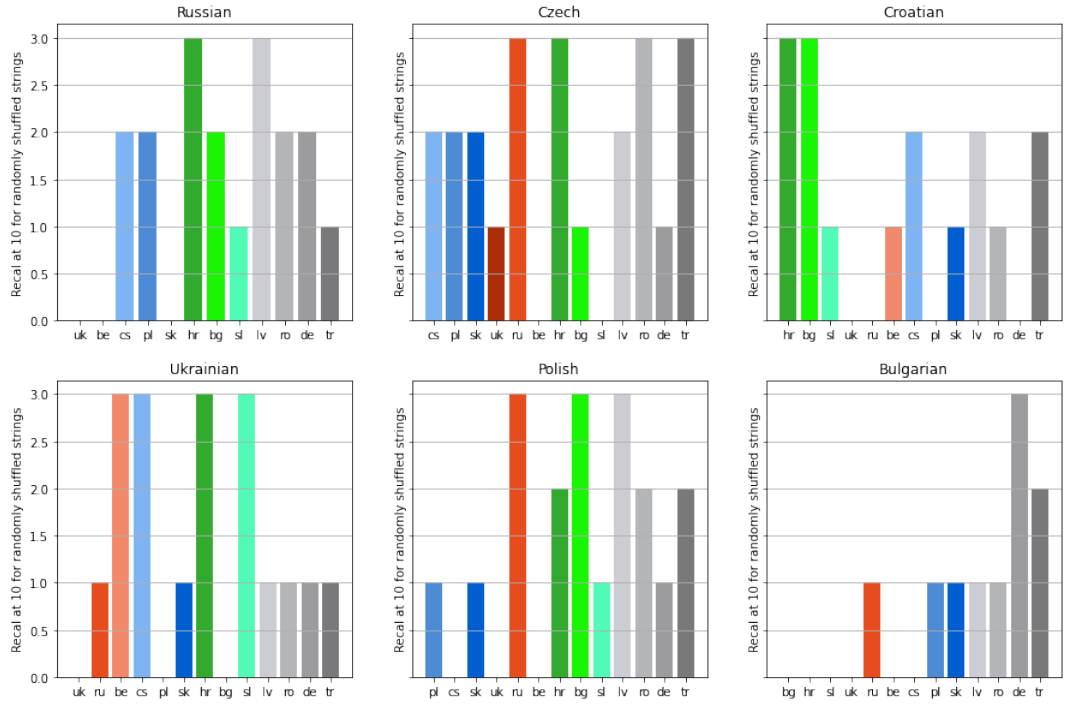


Figure 19: Recall at 10 results for randomly shuffled strings. Each plot corresponds to a model trained on one language. ISO 639-1 codes for the languages: Ukrainian – uk, Russian – ru, Belarusian – be, Czech – cs, Polish – pl, Slovak – sk, Croatian – hr, Bulgarian – bg, Slovene – sl, Latvian – lv, Romanian – ro, German – de, Turkish – tr.

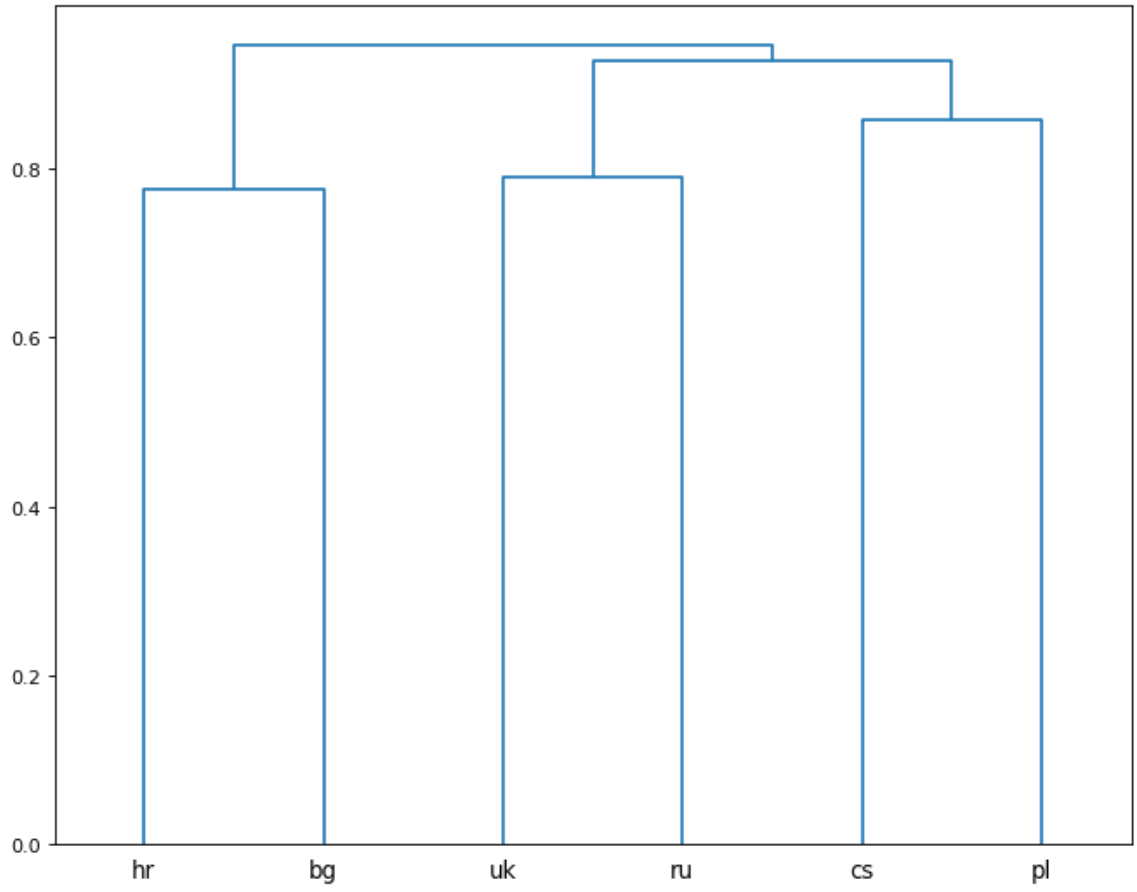


Figure 20: Dendrogram of the Ward clustering of R@10 results. ISO 639-1 codes for the languages: Ukrainian – uk, Russian – ru, Czech – cs, Polish – pl, Croatian – hr, Bulgarian – bg.

Table 3: P-value estimated with bootstrap test on model's R@10. There is a significant difference between most pairs of languages

Language pair	Czech	Polish	Ukrainian	Russian	Bulgarian	Croatian
cs-pl	-	-	0.001	0.435	0.001	0.0009
cs-uk	-	0.237	-	0.0005	0.0001	0.417
cs-ru	-	0.089	0.0	-	0.021	0.301
cs-bg	-	0.084	0.27	0.022	-	0.0
cs-hr	-	0.0005	0.315	0.026	0.0005	-
pl-uk	0.254	-	-	0.0005	0.198	0.001
pl-ru	0.034	-	0.107	-	0.0	0.004
pl-bg	0.008	-	0.0001	0.014	-	0.0001
pl-hr	0.012	-	0.004	0.017	0.0	-
uk-ru	0.007	0.022	-	-	0.0	0.37
uk-bg	0.001	0.017		0.02	-	0.0
uk-hr	0.002	0.0005	-	0.016	0.0	-
ru-bg	0.28	0.475	0.0001	-	-	0.0001
ru-hr	0.343	0.022	0.0005	-	0.097	-
bg-hr	0.423	0.027	0.146	0.476	-	-

all models under analysis are reported in Table 3.

4.3 Quantitative Correlation Analysis

In this subsection, we aim to investigate which characteristics of the input make the model behave as it does. We calculated several types of linguistic distances on the input data to compare the test data coming from different languages. The distances that we used are:

1. Levenshtein Distance (LD), which is a measure of how different two strings are. The difference is calculated as the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

2. Phonologically Weighted Levenshtein Distance (PWLD), which is a measure of phonological similarity between different phonemic sequences or wordforms [7]. The PWLD metric is an extension of the string-based Levenshtein distance that also calculates the cost of each phone substitution based on phoneme features. 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 yxo /u x ɔ/, where phonemes /o/ and /ɔ/ are very similar to each other.

   Following the approach of Abdullah, B.M. et al. (2021) [2], we adapt the PWLD metric in three ways: (1) we present every phoneme in our inventory as a feature vector based on the PHOIBLE [24] feature set, (2) we compute the substitution cost between phonemes as the Hamming distance between their corresponding feature vector, and (3) we set the deletion and insertion cost to 0.5 that equals the

Table 4: Pearson correlation coefficient for metrics under analysis. The retrieval metrics are highlighted with blue, and distance metrics with red

	R@10	MRR	cos sim	LD	PWLD	word len	n cons	n vowels
R@10		0.98***	0.5***	-0.74***	-0.57***	-0.4***	-0.54***	-0.06
MRR			0.5***	-0.75***	-0.56***	-0.41***	-0.56***	-0.08
cos sim				-0.29*	-0.44***	-0.19	-0.21	-0.04
LD					0.8***	0.75***	0.56***	0.38***
PWLD						0.75***	0.7***	0.47***
word len							0.48***	0.31*
n cons								0.31**
n vowels								

If a p-value < 0.05, it is flagged with one star (*). If a p-value < 0.01, it is flagged with 2 stars (**). If a p-value < 0.001, it is flagged with three stars (***)

maximum possible substitution cost.

3. Word length, which, together with the next two metrics, could be seen as another measure of word similarity.

4. Number of consonants, that could show an effect if the model is especially sensitive to consonant sounds.

5. Number of vowels, similarly, would show how sensitive the model is to a particular group of phonemes and help to explain the model's behaviour.

We correlated these distances with our model's retrieval metrics (R@10, MRR, and Cosine Similarity) using Pearson correlation coefficient. The table 4 demonstrates the correlation scores of all the metrics under analysis.

The plots of R@10 and Levenshtein Distance correlation, as well as MRR and Levenshtein Distance correlation are shown in Figures 21 and 22. Correlation of R@10 and PWLD, and correlation of MRR and PWLD are shown in Figures 23 and 24. On all the plots, the pink line represents linear regression on the data.

4.4 Qualitative Analysis

To look at how the retrieved spoken word forms compare to each other and get an insight into the model's meaning recognition logic, we visualized distances of concept retrieval of all models using t-SNE. For t-SNE computation, we used output vectors for all the test data, but visualized only the concepts FOG, WIND, FISH, and WATER. The visualizations are shown in Figure 25 for Bulgarian model, Figure 26 for Croatian, Figure 27 for Russian, Figure 28 for Ukrainian, Figure 29 for Czech, and Figure 30 for Polish. As expected, most of the more similar sounding word forms appear closer to each other on the visualization.

To analyze the results more closely, we provide the top retrieved words for the model trained on Russian and tested on Ukrainian. Table 4.4 shows the top scored pairs as

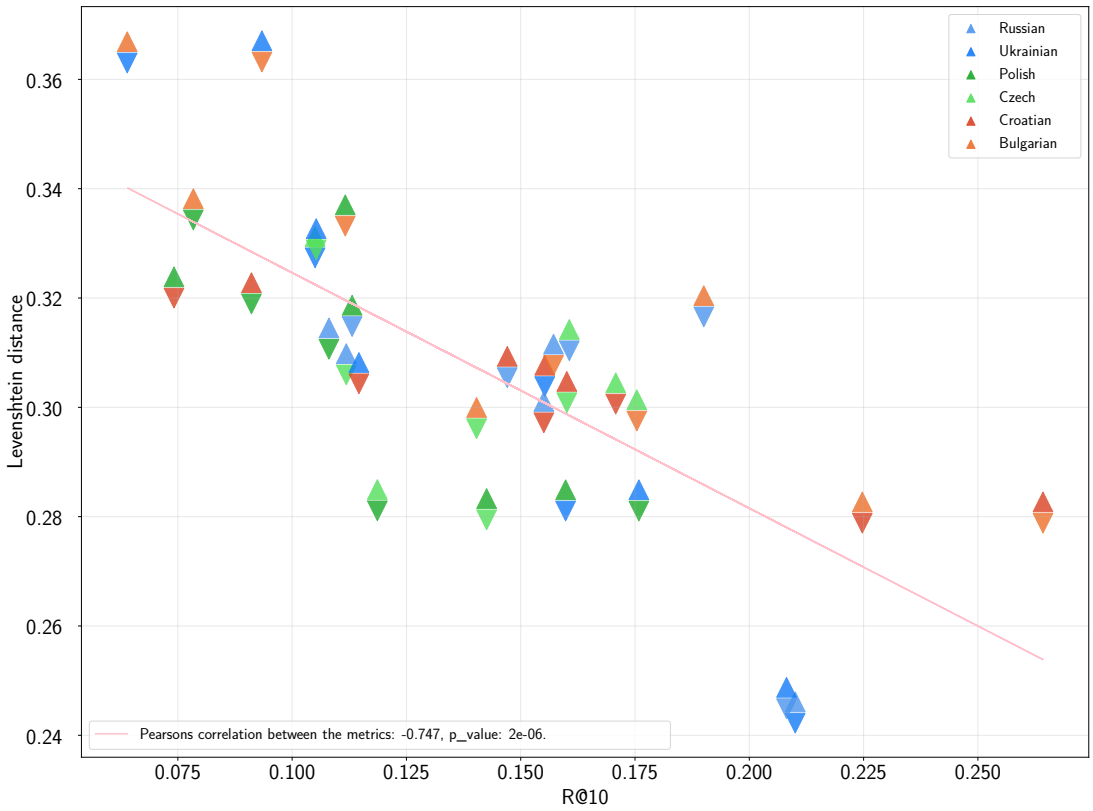


Figure 21: Correlation of R@10 and Levenshtein Distance. The top triangle shows the language of the model, and the bottom triangle shows the test language

Top Cosine Similarity pairs (Russian-Ukranian)	Cosine similarity
/j a/ - /j ɑ/	3.742
/r ɑ n a/ - /r ɑ n̪ ɑ/	1.453
/k t o/ - /x t̪ ɔ/	1.440
/k ɑ ʃʸ a/ - /k ɑ ʃ ɑ/	1.275
/s u p/ - /s̪ u p/	1.242

measured by Cosine Similarity.

The table 5 demonstrates other top scored candidates in Ukrainian for the same phonemic sequences in Russian as in table 4.4. The English translation of the concept is given in the brackets.

Table 5: 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 ʃʸ a/ ('porridge')**	**/s u p/ ('soup')**
Nearest neighbors	- /j ɑ/ ('I') - /d̪ ɛ/ ('yes') - /s̪ i m/ ('if') - /x t̪ ɔ/ ('who') - /j i ʒ ɑ/ ('life')	- /r ɑ n̪ ɑ/ ('wound') - /j ɑ/ ('I') - /ʃ ɑ p k ɑ/ ('hat') - /j i ʒ ɑ/ ('life') - /d̪ ɛ/ ('yes')	- /x t̪ ɔ/ ('who') - /t̪ u t̪/ ('here') - /v ɔ r ɔ ɦ/ ('whisper') - /tʃ ɔ m u/ ('why') - /b i j/ ('was')	- k ɑ ʃ ɑ ('porridge') - /r ɑ n̪ ɑ/ ('wound') - /v ɔ r ɔ ɦ/ ('whisper') - /ʃ ɑ p k ɑ/ ('hat') - /k n̪ ɪ ɦ ɑ/ ('book')	- s̪ u p/ ('soup') - /k ɑ ʃ ɑ/ ('porridge') - /d̪ ɛ n̪ʲ/ ('day') - /x ɑ r tʃ/ ('food') - /k rʲ u k/ ('hook')

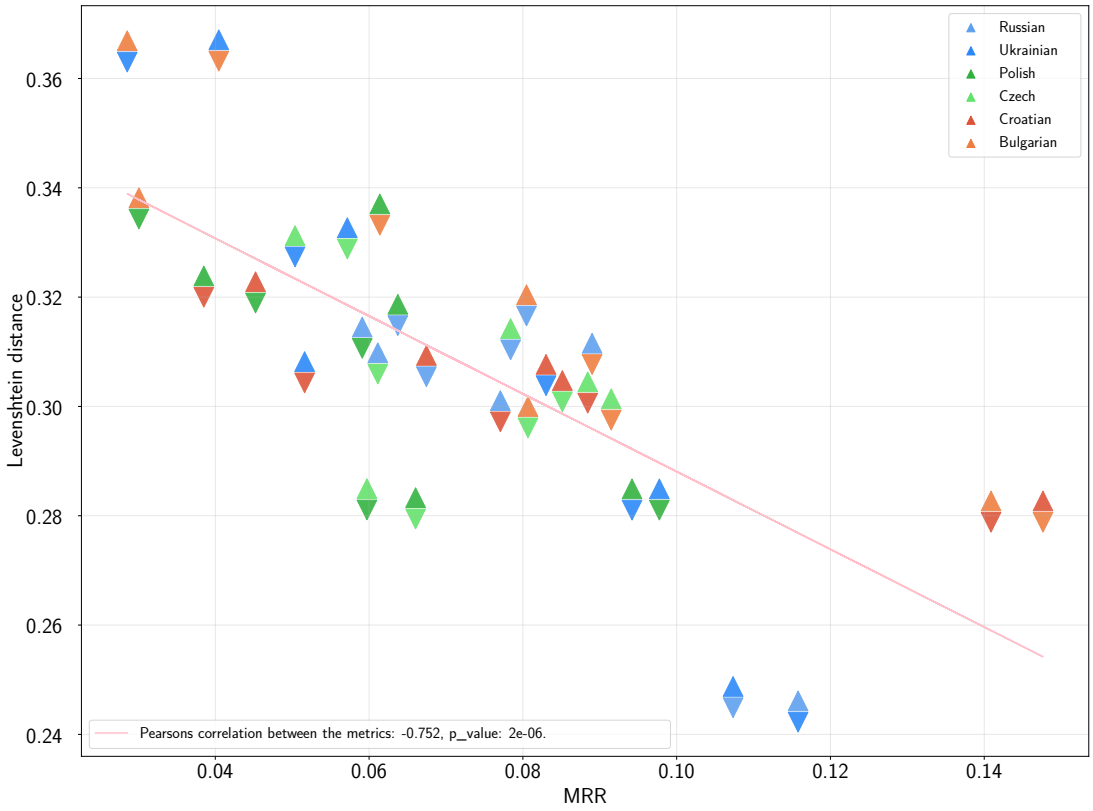


Figure 22: Correlation of MRR and Levenshtein Distance. The top triangle shows the language of the model, and the bottom triangle shows the test language

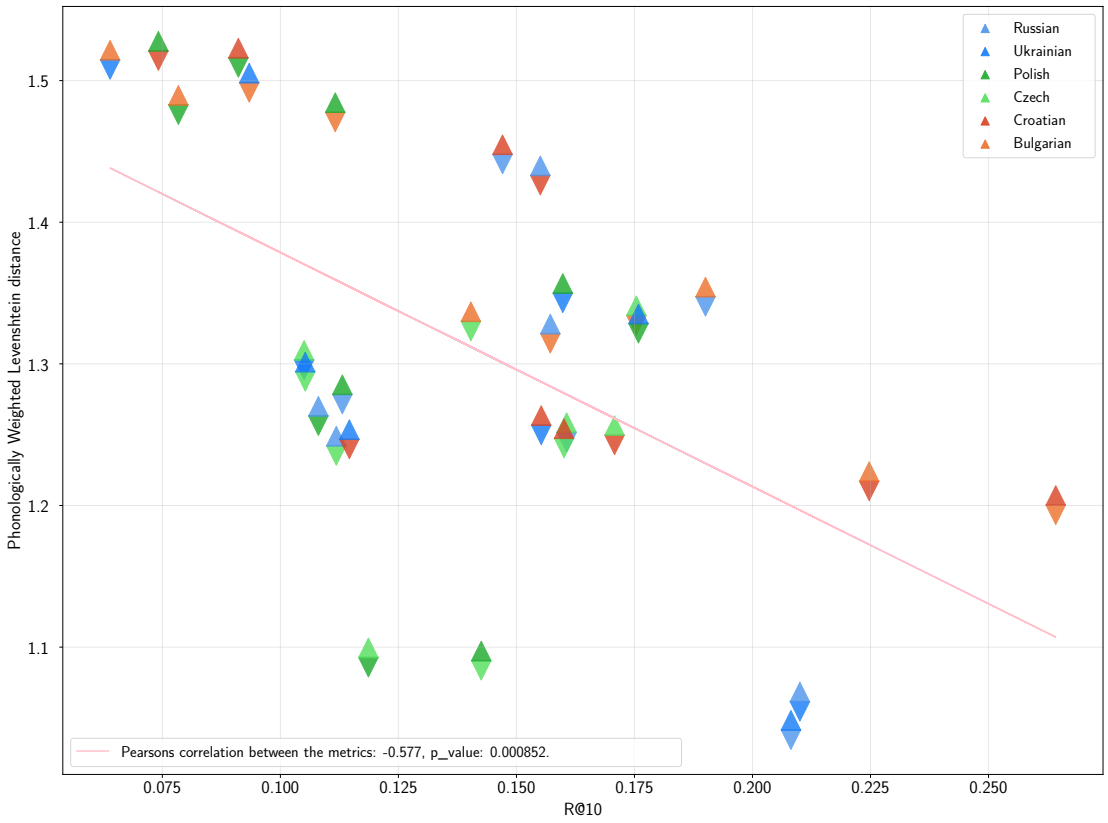


Figure 23: Correlation of R@10 and PWLD. The top triangle shows the language of the model, and the bottom triangle shows the test language

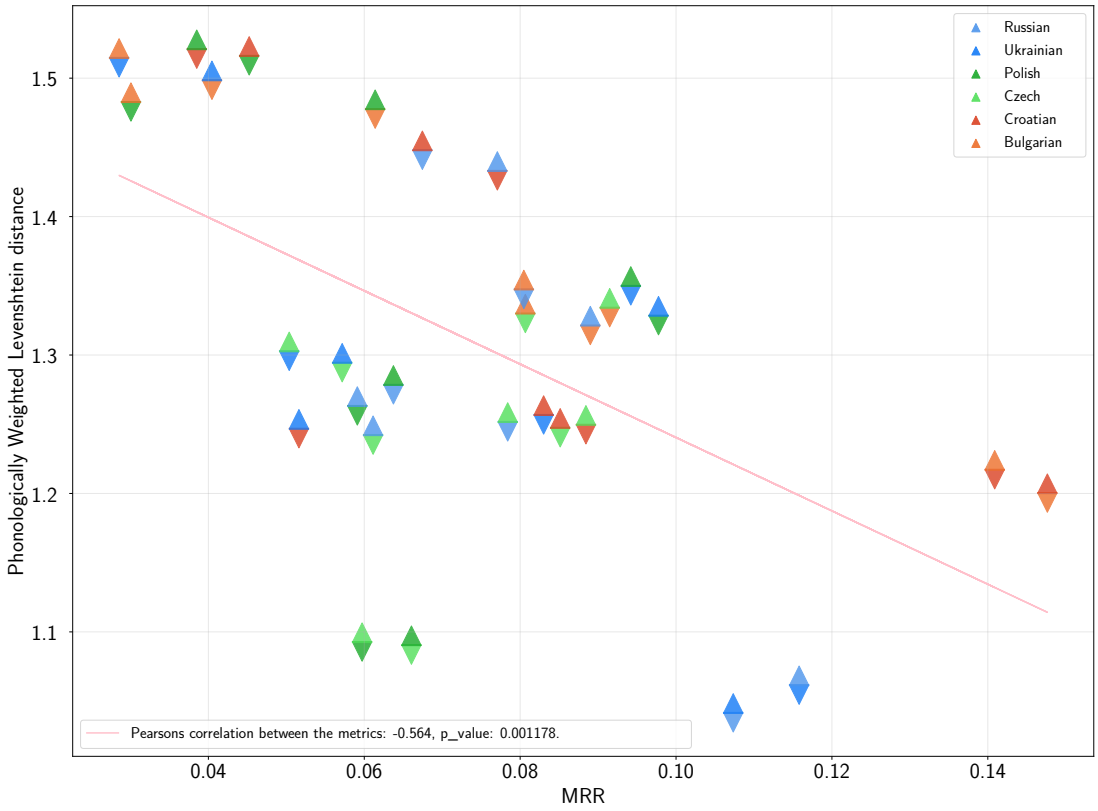


Figure 24: Correlation of MRR and PWLD. The top triangle shows the language of the model, and the bottom triangle shows the test language

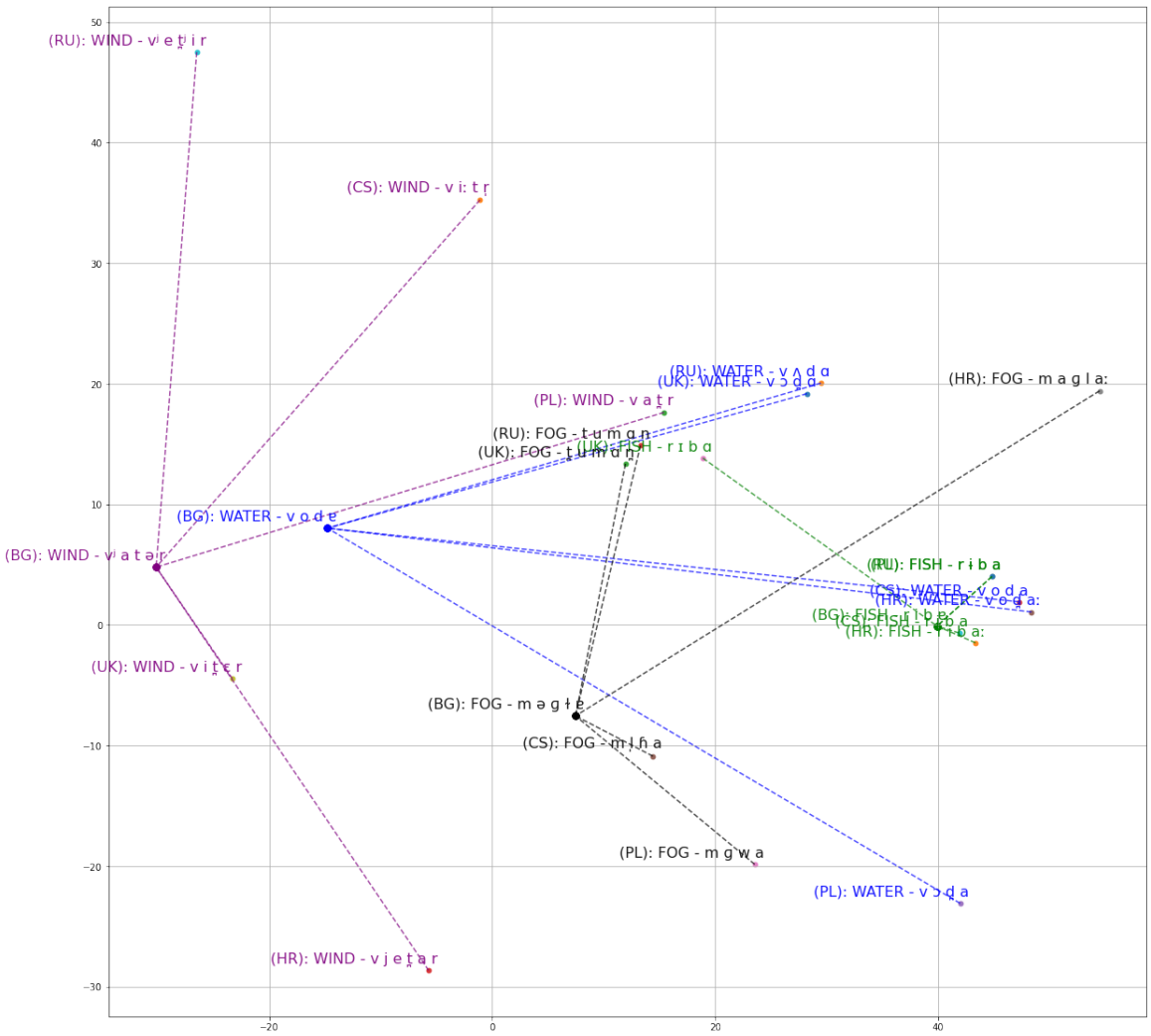


Figure 25: t-SNE on the concept retrieval of the Bulgarian model

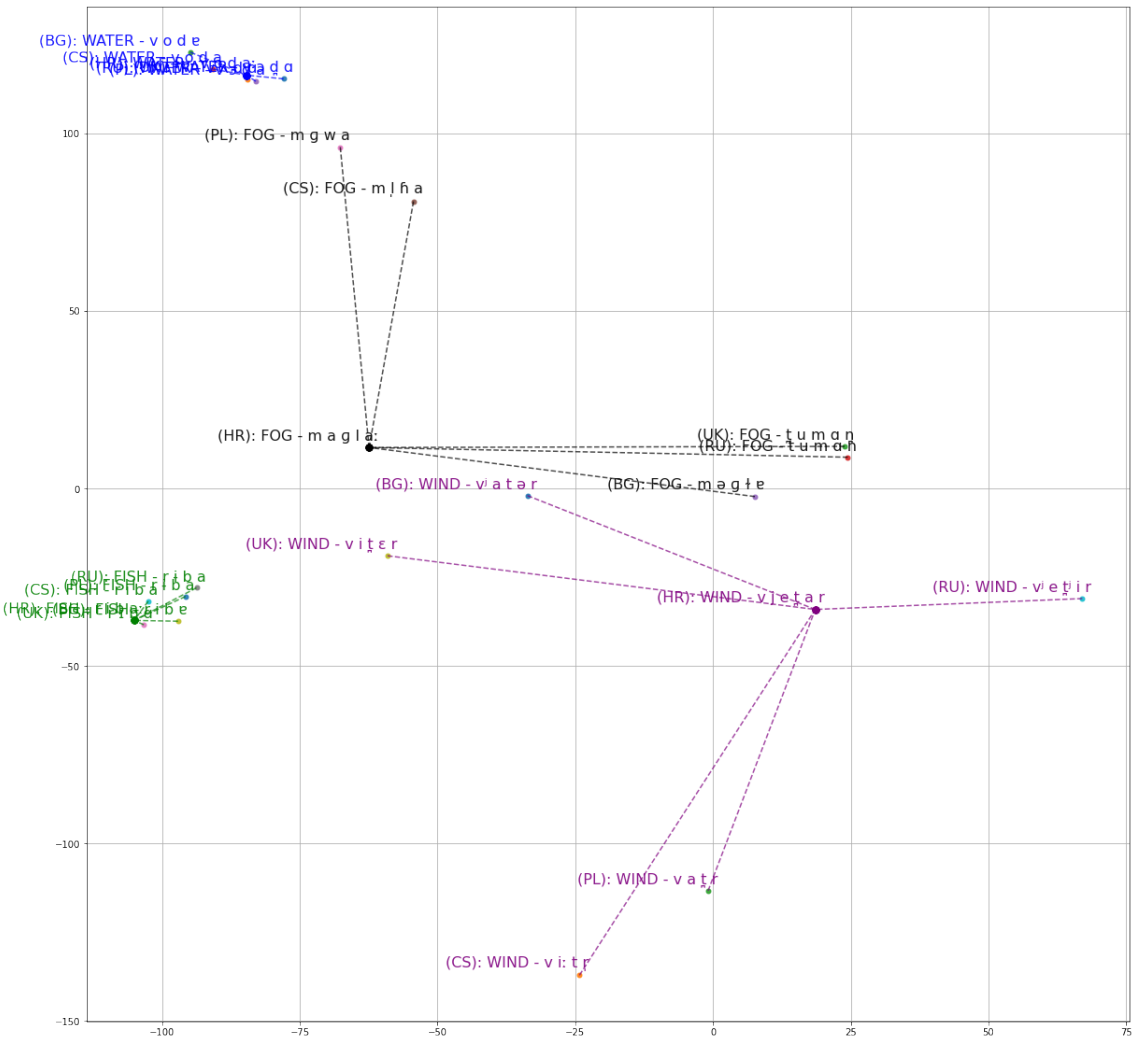


Figure 26: t-SNE on the concept retrieval of the Croatian model

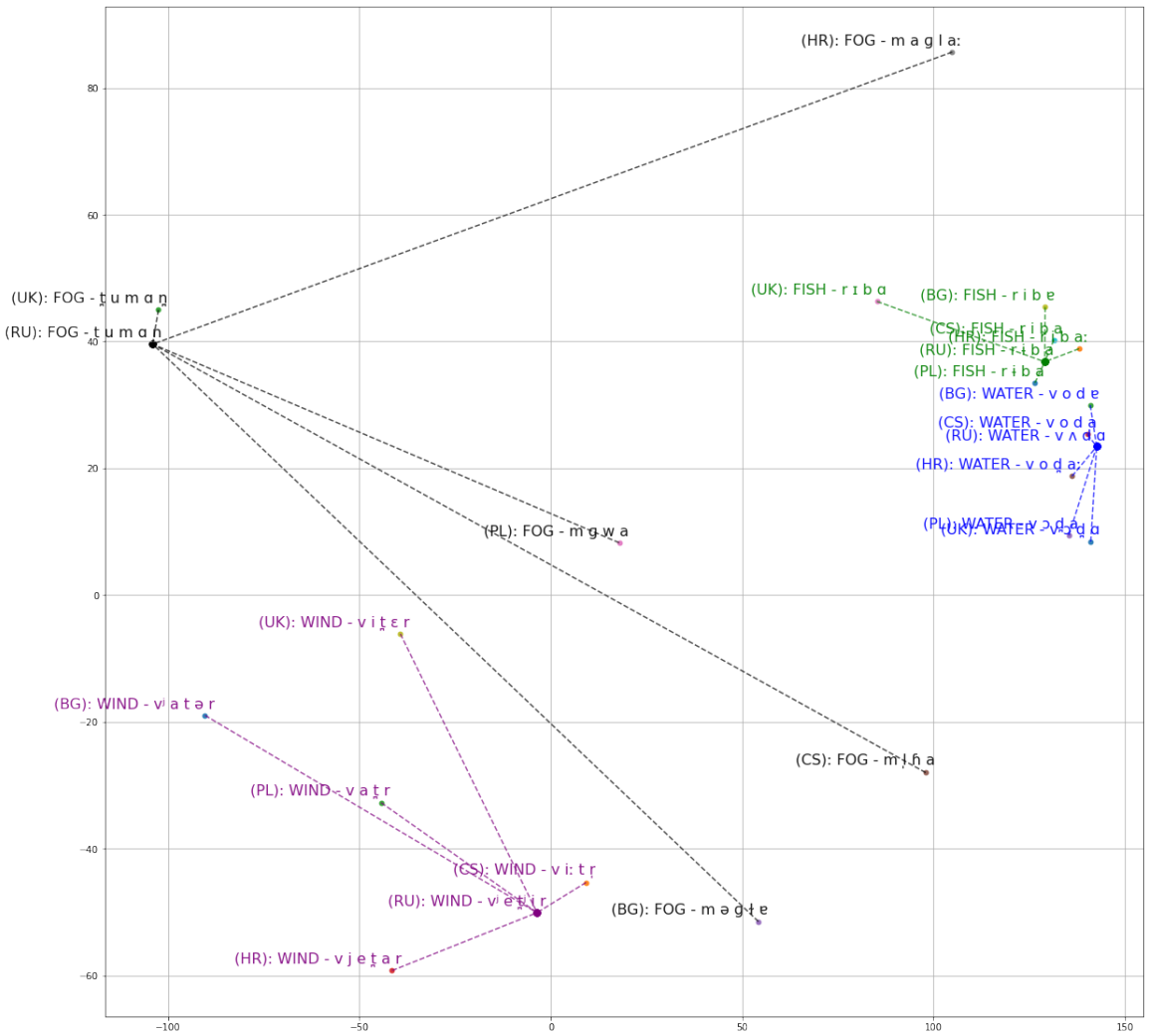


Figure 27: t-SNE on the concept retrieval of the Russian model

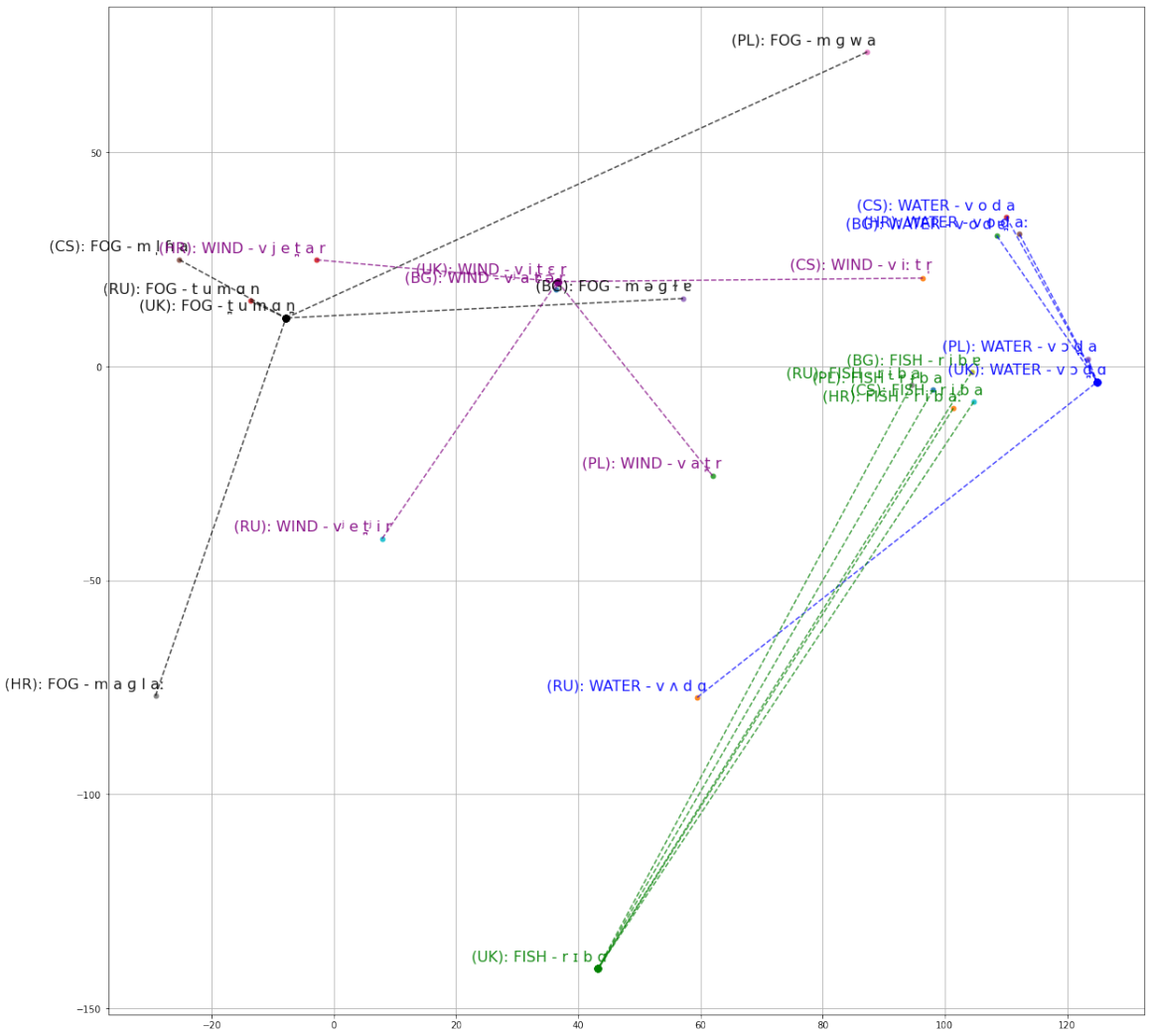


Figure 28: t-SNE on the concept retrieval of the Ukrainian model

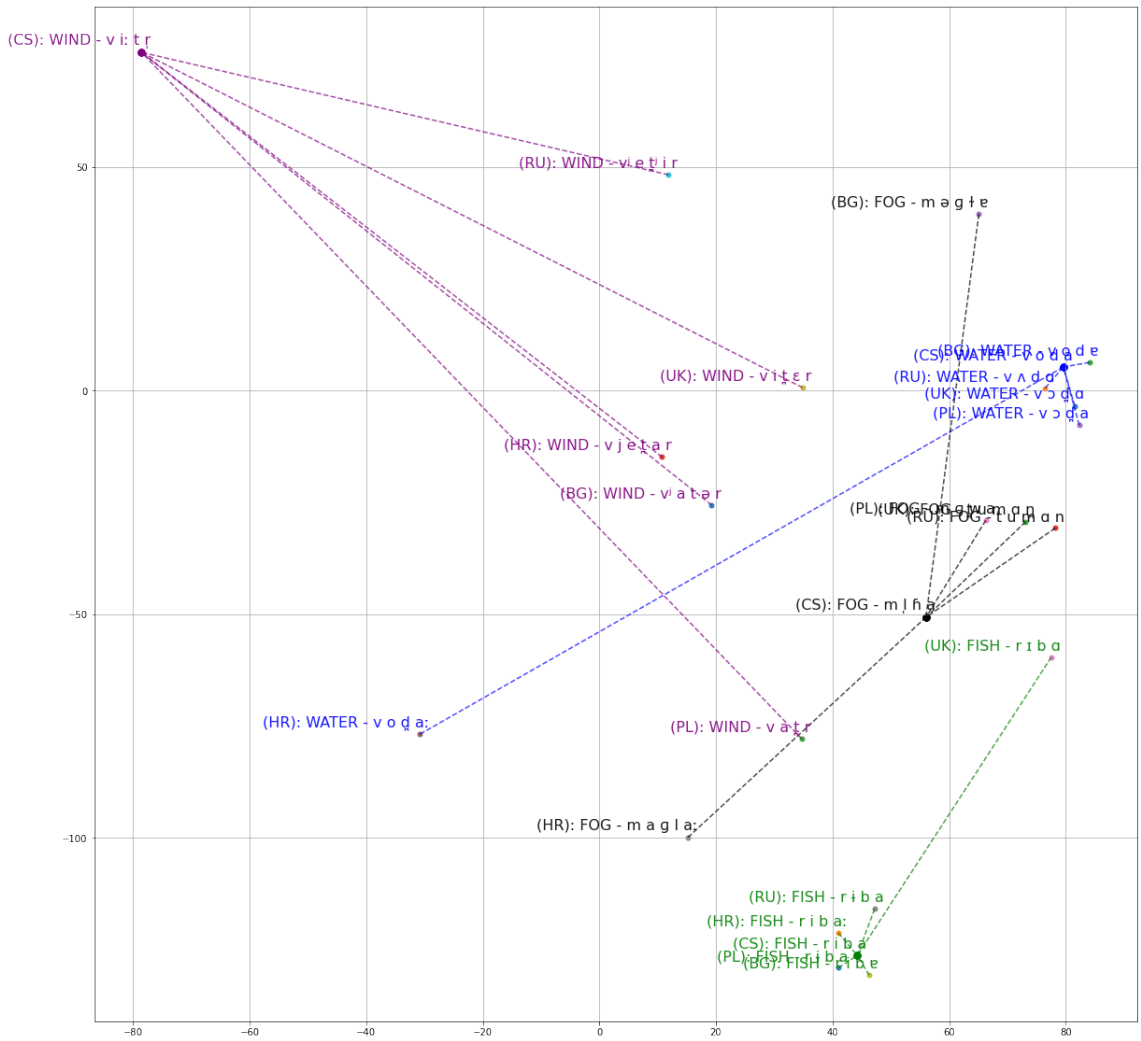


Figure 29: t-SNE on the concept retrieval of the Czech model

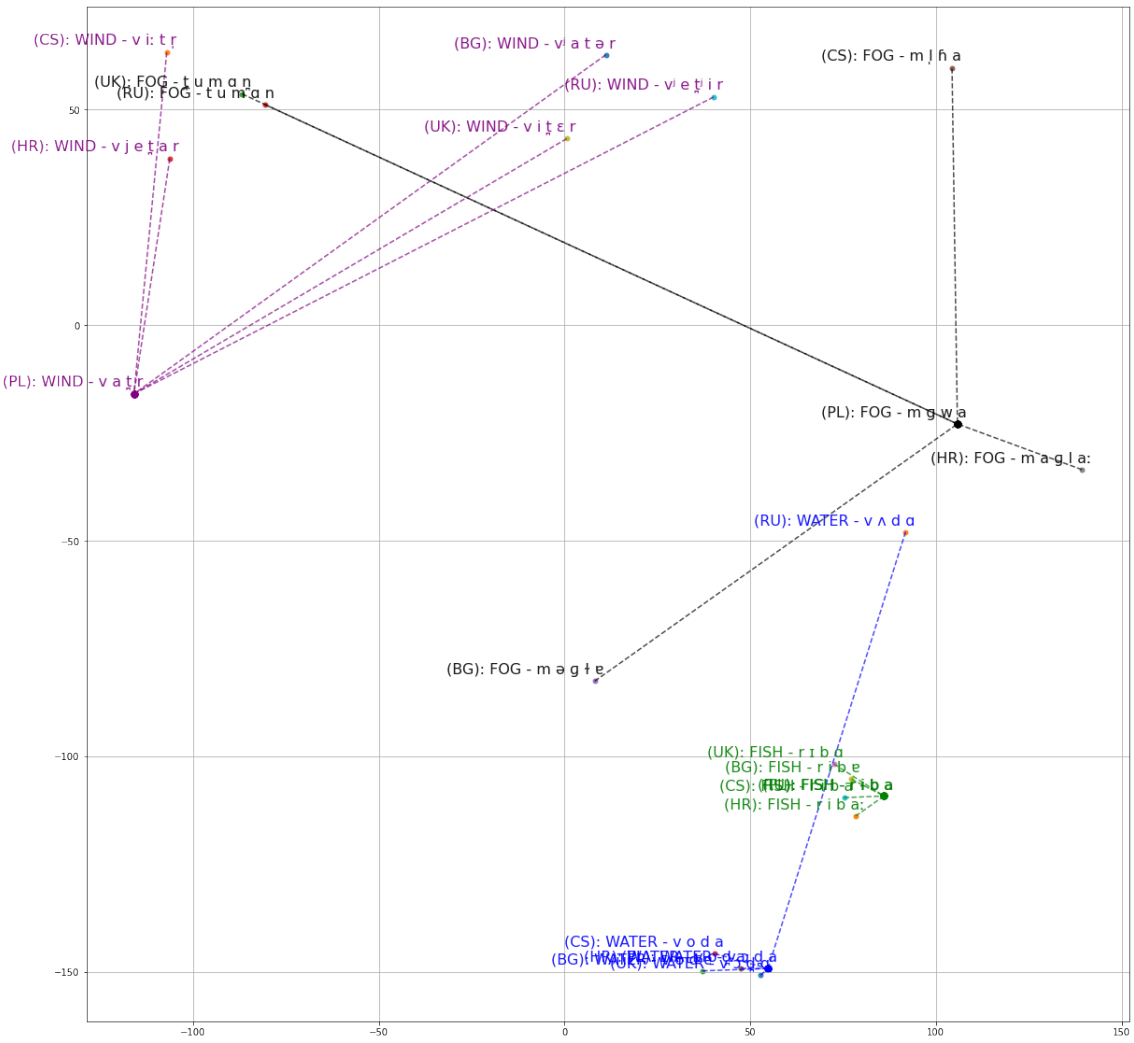


Figure 30: t-SNE on the concept retrieval of the Polish model

5 Analysis of Results

At this point, we have described the structure of our model and data and reported the results of training and testing the model on all the six languages (Russian, Ukrainian, Czech, Polish, Croatian, and Bulgarian). In the previous chapter, we also described some statistical and qualitative experiments we conducted. In the current section, we discuss these experiments and draw conclusions based upon their results.

5.1 Analysis of Model Performance

The main objective of this work is to explore whether the extent to which the model which has only been exposed to one of the 6 Slavic languages under analysis will be able to recognize speech in the other 5 languages. Below, we discuss the performance of the model in line with this goal.

5.1.1 Monolingual Performance

As the first step in assessing the model's recognition of unseen spoken wordforms, we need to set a gold standard for word retrieval and evaluate the model on the language it was trained on.

Figure 4 shows the recall at 1, 5, and 10, as well as mean reciprocal rank scores obtained by the models for all the six languages on monolingual meaning retrieval. As can be seen from the plot, the models produced moderately high scores that are comparable to previous speech and word recognition research [20], [18].

Interestingly, the 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 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). Finding the exact cause of this uniformity would require future research on training the proposed model on languages with different, perhaps more analytic, structure.

5.1.2 Cross-lingual Performance

Having established a benchmark by testing the models monolingually, we can now discuss cross-lingual intelligibility. On the word retrieval score plots for all languages (presented on the Figures 15, 16, 17, 18, we can see that the phonemic sequences in the language from the same subgroup of Slavic languages (such as, Ukrainian and Belarusian for Russian and Croatian and Slovene for Bulgarian) are recognised significantly better than others by most models, which is reflected in all the four sets of scores. Additionally, the retrieval on non-Slavic languages (Latvian, Romanian, German, and Turkish) is

generally visibly lower. This satisfies our hypothesis that the languages which are more genetically related are also more inter-intelligible within the proposed model.

5.1.3 Outliers

The models which retrieval scores raise most questions are the ones trained on West Slavic languages (Czech and Polish). It is much expected that Slovak is recognized very well by the Czech model, since a high intelligibility between the two languages is also confirmed by the results of Golubović, J. and Gooskens, C. (2015)'s experiments [10] (see Figures 3, 5). Scores for Polish, which is genetically very close to Czech, are, on the other hand, one of the worst among all Slavic languages. It is also not clear why the Czech model appears to be that good at recognizing South Slavic languages and the Russian language. One example of a similar case is the results of the written cloze test conducted by Golubović, J. and Gooskens, C. (2015) [10]. The cloze test is a task where a certain number of words are omitted from a text and replaced by a gap. From Table 4 we can see that Croatian is understood better than Polish by native speakers of Czech. Moreover, the participants of Croatian and Slovene can understand both Czech and Slovak (West Slavic languages) better than they can understand Bulgarian, which is also a South Slavic language.

A comparable behavior is observed on the model trained on Polish, where Czech and Slovak have a worse recall and MRR scores than Ukrainian and Belarusian. This could be explained by the fact that some parts of modern Poland and Ukraine used to be inside the Polish–Lithuanian Commonwealth, and have a history of a long-lasting cultural interaction. According to M. Łesiów (1998) [21], the contact and mutual influence between Polish and Ukrainian were the main source of language enrichment for both languages. While Ukrainian and Belarusian are also considered to have a high mutual intelligibility due to their shared history and geographical proximity, it seems logical that the Polish model is good at understanding the Belarusian language as well.

5.2 Correlation with Distance Metrics

The table 4 represents the correlation scores of all the distance metrics with our model's retrieval metrics (R@10, MRR, and Cosine Similarity) using Pearson correlation coefficient. We can see that a moderately high correlation across two types of metrics is observed between R@10 and Levenshtein Distance, as well as MRR and Levenshtein Distance, while the correlation with Cosine Similarity scores are much lower. Surprisingly, Phonologically Weighted Levenshtein Distance has a lower correlation with the retrieval metrics, even though it uses the same phoneme vectorization scheme as the model. It is also interesting that the retrieval metrics R@10 and MRR have a moderate correlation with the number of consonants, which probably has to do with a statistically large number of consonants in Slavic languages, and, consequently, as a high correlation

of this metric with Levenshtein Distance.

5.3 Qualitative Analysis

5.3.1 Top Cosine Similarity pairs and Nearest Neighbors

From the lists of cross-lingual nearest neighbors reported in Table 5, 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 be related to be the nature of fastText embeddings [23] that we used as target embeddings for the model. As already mentioned, fastText is trained with CBOW model, that is able to capture the meaning of a word by combining distributed representations of surrounding words. This way, 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 closely words appear to have a lot in common and are recognized as semantically similar.

Another discovery from Table 5, as well as from Table 4.4, is a clear advantage of shorter and non-content spoken word forms over longer ones, which seems similar to the Word Length Effect, described by S.R. Schmidt [33]. On immediate-recall tests with humans, recall of short words often exceeds recall of long words. There exist several interpretations of the word-length effect in the literature, however, one explanation suggested by Hulme et al. (2004) argues that the word-length effect is in reality an effect of item complexity [13]. Short items are less complex and contain fewer features than long items, and, thus, they will share fewer features across other list items. An application of this hypothesis on our work would be related again to the fastText embeddings: 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.

Finally, the top closest concept pairs appear to be represented by different, but similarly sounding phonemes (for instance, /k t o/ - /x t̪ ɔ/, /r ɑ n a/ - /r ɑ n̪ ɑ). This could mean that the phoneme embeddings that we created using PHOIBLE dataset, that characterizes each phonemes with 38 different features, facilitate the model's understanding of phonemes' similarity. Another factor to support this claim is the advantage of PHOIBLE embeddings over other embeddings that we tried to use, as shown on Table 2.

5.3.2 t-SNE Plots

The t-SNE visualizations are shown in Figure 25 for Bulgarian model, Figure 26 for Croatian, Figure 27 for Russian, Figure 28 for Ukrainian, Figure 29 for Czech, and Figure 30 for Polish. As already said in the previous chapter, words that are phonologically more similar appear closer to each other on the visualization plot.

Most of the time, clear clusters of concepts form, especially if the words sound similar. Again, this probably has to do with the nature of the target fastText embeddings, which are trained to remember the word's context. There are some exceptions to that, such as the concept FOG in the output of Polish model 30. In this case, t-SNE clustered the concept in different languages quite far from each other.

It is interesting that words that do not sound similar, but have a very high semantic proximity, many times appear very close to each other. An example of this are the concepts FISH and WATER, that are quite close on all the plots.

Occasionally, wordforms that are not semantically connected to each other and do not sound similar also appear in the same cluster (for instance, in Figure 28, the Russian, Czech, and Ukrainian words for FOG are clustered with the Croatian concept for WIND. A reason for that could be the semantic proximity of the given concepts, i.e. all of them semantically are sense organs, or simply the model failing to generalize on certain wordforms. In this case, we can conclude that the model is not always successful in differentiating among words from related semantic groups.

At times, it is hard to interpret the results of clustering, since to our perception, many similar sounding words are not recognized as such by the model (for instance, the concepts WIND and WATER in the output of the Bulgarian model in Figure 25). Our hypothesis is that there could be other (unexplored) factors that cause this behaviour.

6 Conclusion and Future Work

In this work, we presented a spoken-word recognition model based on a multi-layer Long Short-Term Memory (LSTM) neural network that maps variable-length phonemic representations of wordforms into their meaning representations and simulates cross-lingual lexical processing. The model is trained to recognize the meanings of spoken wordforms for each of six Slavic languages under analysis (Bulgarian, Croatian, Czech, Polish, Russian, and Ukrainian). On the abstract level, we aim to model a listener, that has only heard a singular language in their lifetime.

Our primary research goal is to computationally test to what extent a computational model which has only been exposed to one language would be able to recognize the meaning of spoken words in closely related languages. In general, our model manages to simulate the results of previous sociolinguistic studies on language intercomprehension [10]. We obtain quite encouraging results for both mono- and cross-lingual lexical retrieval, that are comparable to previous work on computational modeling of speech recognition [18], [20].

6.1 Contributions

The contributions this work presents can be summarized in the following points:

1. We approximately simulate language relatedness between Slavic languages using computational modeling. According to our initial hypothesis, the results for most training languages reflect the official genealogical classification of languages.
2. Phoneme embeddings created using PHOIBLE features significantly improve a neural network model's understanding of similarity between phonetic sequences. It is especially relevant in the context of multilingual language modeling, where a large number of allophones of the same phoneme could be encountered.
3. We investigate the model's behaviour by studying the relationship between the input and the results of model's performance. We conclude that the model is sensitive to such linguistic distances as Levenshtein Distance and Phonologically Weighted Levenshtein Distance (described in the Chapter 4), as well as to word length and number of consonants.
4. Qualitative analysis of the model's output sheds some light on its flaws. The t-SNE visualizations demonstrate that occasionally the model is not able to cluster similarly sounding words together.

6.2 Future Work

For future work, there are several possible research directions that could be explored:

1. An expansion of this project to include a more complex model and possibly improve word retrieval scores. Using attention mechanism, for example, could be beneficial to improve the model's semantic generalization.

2. Another idea that could possible help the model to learn the structure of the learned wordforms is inclusion of various linguistic features, such as lemmas, syntactic dependency labels, and morphological features, into the data. Even though we are trying to model a human listener, it is possible that humans learns these features explicitly by exposure to the word on the sentence and text level.

3. Although we have found a strong correlation of our model's output with certain linguistic distances, we hypothesize that there might be other factors that cause the model to behave the way it does. Potentially, one could experiment with other input metrics (e.g., token to type ratio, number of cognates in parallel test data, etc.) and investigate the influence of inflections to semantic generalization.

4. An extension of this work to language from other language groups would be useful for further explanation of model's behaviour. One could experiment with languages that are less inflectional than Slavic languages and see whether that makes an effect on training and testing the model.

5. Finally, it would be interesting to compare our results with the human performance on the same data to see to what extent our work models human language learning.

List of Figures

List of Tables

References

[1] Available: http://espeak.sourceforge.net/index.html, accessed: 28.04.2022.

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

[3] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot: Distributed Word Representations for Multilingual NLP. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[4] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An Empirical Investigation of Statistical Significance in NLP. 2012.

[5] Gerda J. Blees and Jan D. ten Thije. Receptive Multilingualism and Awareness. pages 1–13, 2016.

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

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

[8] M. Gareth Gaskell and William D. Marslen-Wilson. Integrating form and meaning: A distributed model of speech perception. Language and cognitive Processes, 12(5-6):613–656, 1997.

[9] Jelena Golubovic. Mutual intelligibility in the Slavic language area. Groningen: Center for Language and Cognition, 2016.

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

[11] Charlotte Gooskens, Vincent J. van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt. Mutual intelligibility between closely related languages in Europe. International Journal of Multilingualism, 15(2):169–193, 2018.

[12] Florian Hintz, Suzanne R. Jongman, Marjolijn Dijkhuis, Vera van ‘t Hoff, James M. McQueen, and Antje S. Meyer. Shared Lexical Access Processes in Speaking and Listening? An Individual Differences study. Journal of Experimental Psychology: Learning, Memory, and Cognition, 46(6):1048–1063, 2020.

[13] Charles Hulme, Aimée M Suprenant, Tamra J Bireta, George Stuart, and Ian Neath. Abolishing the word-length effect. Number 30. Journal of Experimental Psychology: Learning, Memory, Cognition, 2004.

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

[15] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations, 12 2014.

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

[17] Andrew L. Maas, Stephen D. Miller, Tyler M. O'Neil, and Andrew Y. Ng. Word-level Acoustic Modeling with Convolutional Vector Regression. 2012.

[18] Nicole Macher, Badr M. Abdullah, Harm Brouwer, and Dietrich Klakow. Do we read what we hear? modeling orthographic influences on spoken word recognition. In EACL, 2021.

[19] William D. Marslen-Wilson and Alan Welsh. Processing interactions and lexical access during word recognition in continuous speech. Cognitive psychology, 10(1):29–63, 1978.

[20] Alexandra Mayn, Badr M. Abdullah, and Dietrich Klakow. Familiar words but strange voices: Modelling the influence of speech variability on word recognition. In EACL, 2021.

[21] Robert De Lossa MICHAŁ ŁESIÓW and Roman Koropeckyj. The Polish and Ukrainian Languages: A Mutually Beneficial Relationship.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. Curran Associates, Inc., 26, 2013.

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

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

[25] David R. Mortensen, Siddharth Dalmia, and Patrick Littell. Epitran: Precision G2P for many languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, Proceedings of the Eleventh Inter-

national Conference on Language Resources and Evaluation (LREC 2018), Paris, France, May 2018. European Language Resources Association (ELRA).

[26] Dennis Norris. Shortlist: a connectionist model of continuous speech recognition. pages 52(3):189–234, 1994.

[27] 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. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.

[28] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[29] Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. Mimicking word embeddings using subword rnns. arXiv preprint arXiv:1707.06961, 2017.

[30] David C. Plaut. Connectionist Modeling of Language: Examples and Implications. Mind, brain, and language: Multidisciplinary perspectives, 2000.

[31] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020.

[32] Douglas L. T. Rohde and David C. Plaut. Connectionist Models of Language Processing. Cognitive Studies, 10(1):10–28, 2003.

[33] Editor-in-Chief: John H. Byrne Schmidt, Stephen R. Learning and Memory: A Comprehensive Reference. 2.09 - Distinctiveness and Memory: A Theoretical and Empirical Review. 2007.

[34] Kristof Strijkers and Albert Costa. Riding the lexical speedway: a critical review on the time course of lexical selection in speech production. Frontiers in psychology, 2:356, 2011.

[35] Roland Sussex and Paul Cubberley. The Slavic Languages. Cambridge University Press, 2006.

[36] Chaoju Tang and Vincent J. van Heuven. Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. Linguistics, 53(2):169–193, 2015.

[37] Jan D. ten Thije and Ludger Zeevaert. Receptive Multilingualism. John Benjamins Publishing Company, 2007.

[38] Anita E. Wagner, Paolo Toffanin, and Deniz Başkent. The Timing and Effort of Lexical Access in Natural and Degraded Speech. Front. Psychol, 7(398), 2016.