Paper Review

Emma Angela Montecchiari
[emma.montecchiari@studenti.unitn.it]

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

I. Zaitova, B.M. Abdullah and D. Klakow, 2022
Proceedings of the Ninth Workshop on NLP for Similar Languages,
Varieties and Dialects, pages 54–63 October 16, 2022.

University of Trento - Cognitive Science Master's degree
Course: Human Language Technologies
Winter Session Submission - February 9, 2024

1 Paper overview

1.1 What is the problem addressed in the paper?

Closely related languages exhibit varying degrees of mutual *intelligibility*, a measure of cross-language similarity. It is a uni-dimensional metric of linguistic distance influenced by many structural linguistic features (morphology, vocabulary, phonetics, etc.) ([24], [8], [11]). Psycholinguistic and sociolinguistic studies explore the effects of intelligibility, such as *intercomprehension*. This is the ability of speakers of intelligible languages of (partially) comprehending each other's speech to a great degree without explicitly learning the second language (L2).

Empirical testing on intelligibility and intercomprehension in psycholinguistics involves studies on lexical surprisal [10], cross-lingual priming [12], word translation, and tasks like the cloze test and picture naming [2].

Key linguistic aspects influencing intelligibility include *lexical distance* driven by cognates, words encoding the same concepts with similar phonological forms across languages. As well as *phonological distance*, measured by metrics like Levenshtein distance, computing the smallest number of string edit operations needed to convert the string of phonetic symbols ([7], [9]).

Spoken-word recognition theories and models aim to explain the process of accessing lexical knowledge from acoustic realization [13]. It can be considered as a lexical access problem, copying with the variability of speech, retrieving phonological and semantic information [22].

In the cognitive modeling framework, 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. Valid computational frameworks have been shown to resemble human behaviour in speech tasks ([14], [17], [16], [3], [13]).

Computational approaches prove useful for testing hypotheses and isolating linguistic levels' effects on language processing. The authors present a neural model of spoken-word recognition investigating the degree to which a monolingual model, i.e. trained on a single language, is able to recognize the meaning of spoken words across related languages.

1.2 What are the stated objectives of the work presented in the paper?

The study explores mutual intelligibility, focusing on Slavic languages, known for typological and structural similarities ([23], [6]). Genetic proximity influences cross-language intelligibility, with shared ancestry enhancing mutual understanding ([7], [4]). Slavic languages, exemplifying high mutual intelligibility, show intra-family dynamics, revealing greater understanding within the same sub-family ([10], [12], [20]).

Methodologically, a recurrent neural network, optimized with ADAM and MSE loss, is proposed. The model comprises one LSTM layer followed by a linear-tanh MLP. Phonological sequences input is transformed using PHOIBLE feature set [1], and FastText embeddings in a CBOW algorithm populate the semantic space [18].

Six monolingual models have been trained on Russian, Ukrainian, Polish, Czech, Bulgarian, and Croatian. Cross-lingual evaluation spans Slavic (Belarusian, Slovak, Slovene) and non-Slavic languages (German, Romanian, Latvian, Turkish).

Objectives include predicting mutual intelligibility, reflecting genetic relations, and identifying linguistic measures correlated with cross-lingual performance. The study delves into the correlations between linguistic mutual intelligibility and computational modeling.

2 Paper overcomes

The model undergoes training on monolingual data, sampling word forms from FastText embeddings. Concepts, shared across six languages, are meticulously chosen for linguistic uniformity. Both monolingual and cross-lingual evaluations utilize FastText embeddings from the same training space.

During testing, the model computes the meaning representation of the phonemic sequence in the test language. To evaluate model retrieval on the test set, cosine similarity finds the closest match between the model output and the target vector from the model training language. Cosine similarity is then calculated against all possible ground truth vector representations in the language of training, encompassing both monolingual and cross-lingual settings.

Performance metrics, including average recall at 1, 5, and 10, along with Mean Reciprocal Rank (MRR), are computed for both monolingual and cross-lingual evaluations. In cross-lingual evaluation, the cosine MRR similarity of the L2 target concept is measured against all evaluation concepts in the embedding space of the model training language.

Linguistic predictors are scrutinized through Pearson correlation, examining phonetic-lexical distances like Levenshtein Distance and Phonologically Weighted Levenshtein Distance (PWLD) [5]. Hierarchical clustering analyzes Recall@10 results among the six models, contributing to a comprehensive understanding of the model's performance.

2.1 What ideas presented in the paper worked well? What are the key contributions of this paper?

In evaluating Slavic languages, the model demonstrates significantly better recognition for phonemic sequences within the same sub-group (e.g., Ukrainian for Russian, Croatian for Bulgarian). Performance on non-Slavic test forms (Latvian, Romanian, German, and Turkish) is generally lower, except for Bulgarian. Cross-linguistic phonetic-lexical similarities correlate with the model's concept retrieval performance, aligning with sociolinguistic observations. Cluster analysis accurately reconstructs the genealogical Slavic language tree, showcasing the model's ability in this aspect.

The model's architecture, utilizing LSTM and MLP, is chosen for its proven cognitive validity in predicting human behavior and cognitive features. LSTM, especially, mirrors the cognitive aspects of lexical retrieval and addresses the gradient problem, ensuring robust implementation. Additionally, the model adheres to the multiple-trace theory, departing from traditional word recognition and lexical access ([21], [15], [19]). Embracing contemporary mental lexicon approaches, it posits multiple entries for each word in lexical memory, in the form of detailed perceptual traces that preserve fine phonetic detail of the original articulatory event. Episodic approaches highlight the continuity and close coupling between speech perception, production, and memory in current language processing theories.

Furthermore, the architecture aligns with a traditional phonological view, segmenting phonemes within a rule system governing sound patterns and sequences (discrete phones, features, allophones). This perspective seamlessly integrates with statistical learning frameworks. This choice adeptly tackles the acoustic-phonetic invariance challenge by identifying critical acoustic attributes across contexts. Unlike traditional speech-word models, it sidesteps the complexities of speech perception variability.

2.2 What ideas turned out not to work very well? What are the weakest parts of this paper?

As for the model cross-lingual performance results, unexplained phenomena are such that Bulgarian recognized Romanian evaluation set better than Ukranian, of the same language family. The Czech model shows unexpected disparities in recognizing Polish word forms and high retrieval performance on Slovak word forms. As well as the Russian model, which exhibits surprising recognition patterns for Croatian and Bulgarian compared to Belarusian. For the Bulgarian case, they propose that the geographic proximity in this case could lead to lexical borrowing.

Moreover, the more precise phonological measure, PWLD, has a lower correlation with retrieval metrics than LD, despite using the same phoneme vectorization scheme as the model. The t-SNE clustering results raise

concerns about the alignment of similarly sounding words in different languages. Non-similar-sounding words like 'mosquito' and 'wind' appear close, possibly due to the nature of FastText embeddings trained to predict word context.

A notable drawback, in my opinion, is the lack of accessibility to implementation code, training/testing data, and outcomes. This limitation hinders the opportunity to explore the performance across specific word categories, which could have provided valuable insights. Additionally, the paper lacks substantial evidence regarding the training of distributional semantic spaces. The absence of detailed information on the training data for these spaces is crucial for a thorough analysis of the results and could significantly impact the interpretation of outcomes. Moreover, the shortcomings in the outcomes are linked to the semantic embeddings of the outputs, with the paper not explicitly addressing the reasons behind these observed limitations.

3 Suggestions

3.1 If you were to re-write this paper, what parts of the paper would you write differently?

The explanation of the transformation from phonological vector representation to the target semantic representation is not sufficiently detailed. Providing additional word outputs as examples would have enhanced clarity.

The inclusion of a visual representation, such as a genealogical tree map for Slavic languages, would have presented a more straightforward and explanatory figure to illustrate the relationship between training, test, and evaluation sets.

The narrative style could be improved for better coherence, adopting a more linear approach and providing a more thorough explanation of the theoretical aspects. The structure appears non-linear with numerous internal references, making it challenging to follow seamlessly.

3.2 Was the paper self-contained? What additional references did you consult to understand the paper?

The functions in the model's architecture have undergone cognitive validation in previous studies, although this crucial aspect is not explicitly mentioned. Providing such information would have added value to the presentation.

Additionally, a deeper historical framework could have been provided, delving into the computational modeling of speech processing. This aspect appears under-addressed, and the transition from continuous, parametric, and gradient information in the speech signal seems theoretically significant. The study could have benefited within this comprehensive framework.

While the bibliography is comprehensive, a careful reading is necessary to fully grasp the paper's contextual framework. The overall synthesis feels a bit concise.

4 Final Remarks

4.1 Rate this paper on a 1-10 scale = 8

4.2 Suggestions on future directions

(A) Utilize a bilingual training set and observe performance variations. Investigate whether multilingualism enhances inter-comprehension abilities, not only within intelligible language systems but also beyond them.

(B) Explore a behavioral paradigm for model testing. In a monolingual modeling framework, consider written semantic priming with the training word set and output words, along with free speech item generation based on the training words set. In a multilingual modeling scenario, attempt tasks like hinting word meanings of other languages through auditory word input, engaging L1 speakers in free generation tasks from L2 words inputs, and conducting word similarity assessments between your language and others captured by the model.

(C) Experiment with languages known for asymmetrical intelligibility, such as Spanish and Portuguese.

References

[1] Phoible 2.0, 2019.

[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] Badr M. Abdullah, Iuliia Zaitova, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. 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, 2021.

[4] Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. A probabilistic generative model of linguistic typology. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] 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)*, page 650, 2016.

[6] Jelena Golubovic and Charlotte Gooskens. Mutual intelligibility between west and south slavic languages. *Russian Linguistics*, 39:351–373, 2015.

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

[8] Charlotte Gooskens. Dialect intelligibility. In *The Handbook of Dialectology*, pages 204–218. 2017.

[9] Charlotte Gooskens. Receptive multilingualism. In *Multidisciplinary Perspectives on Multilingualism*, pages 149–174. 2019.

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

[11] D. Jan and Ludger Zeevaert. *Receptive Multilingualism: Linguistic Analyses, Language Policies, and Didactic Concepts*, volume 6. John Benjamins Publishing, 2007.

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

[13] Nicole Macher, Badr M. Abdullah, Harm Brouwer, and Dietrich Klakow. 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, 2021.

[14] James S. Magnuson, Heejo You, Sahil Luthra, and et al. Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, 44(4):e12823, 2020.

[15] William D. Marslen-Wilson. Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2):71–102, 1987.

[16] Yevgen Matusevych, Herman Kamper, Thomas Schatz, Naomi H. Feldman, and Sharon Goldwater. 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*, 2021.

[17] Alexandra Mayn, Badr M. Abdullah, and Dietrich Klakow. 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, 2021.

[18] 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.

[19] J. Morton. Word recognition. In J. Morton and J.C. Marshall, editors, *Psycholinguistic Series II*. Elek Scientific Books, London, 1979.

[20] Marius Mosbach, Irina Stenger, Tania Avgustinova, and Dietrich Klakow. 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, 2019. INCOMA Ltd.

[21] R. C. Oldfield. Things, words and the brain*. *Quarterly Journal of Experimental Psychology*, 18(4):340–353, 1966.

[22] David R. Pisoni and Susannah V. Levi. Some observations on representations and representational specificity in speech perception and spoken word recognition. In *The Oxford Handbook of Psycholinguistics*, pages 3–18. 2007.

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

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.