Do we read what we hear?
Modeling orthographic influences on spoken word recognition

Nicole Macher   **Badr M. Abdullah**   **Harm Brouwer**   **Dietrich Klakow**

Department of Language Science and Technology (LST), Saarland University, Germany

Corresponding author: `macher.nicole@gmail.com`

Abstract

Theories and models of spoken word recognition aim to explain the process of accessing lexical knowledge given an acoustic realization of a word form. There is consensus that phonological and semantic information is crucial for this process. However, there is accumulating evidence that orthographic information could also have an impact on auditory word recognition. This paper presents two models of spoken word recognition that instantiate different hypotheses regarding the influence of orthography on this process. We show that these models reproduce human-like behavior in different ways and provide testable hypotheses for future research on the source of orthographic effects in spoken word recognition.

1 Introduction

The abstract theory of spoken word recognition (SWR) assumes that the process of speech recognition comprises two phases: a prelexical and a lexical level (Scharenborg and Boves, 2010). The prelexical level contains prelexical representations, like phonological units, which are the result of having processed the raw acoustic signal. These units are assumed to be activated before accessing meaning representations of words in the lexical level. By instantiating the process of SWR in a computational model the underlying theory can then be validated or further refined based on insights into the model's architecture and its behavior.

Influential models of SWR are for example the Cohort model (Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980; Marslen-Wilson, 1987), the TRACE model (McClelland and Elman, 1986) or the Shortlist model (Norris, 1994). These models typically have a connectionist architecture with localist or feature-based representations as their inputs and outputs (Weber and Scharen-

borg, 2012), usually mapping phonological onto semantic representations. There is evidence, however, that orthographic information could be co-activated during phonological processing. For example, words with frequent and consistent sound-spelling relations have been proven to be beneficial for auditory word recognition (*orthographic consistency effect*, initially discovered by Ziegler and Ferrand, 1998). *Consistent* words, i.e., words with phonological rhymes that can be spelled in only one way (e.g. /ʌk/ – *uck*, as in *duck*) produce shorter reaction times in a lexical decision task, thus are easier to process, compared to *inconsistent* words whose rhymes can be spelled in multiple ways (e.g. /aɪp/ can be spelled *ipe* like in *pipe* or *ype* like in *type*). This effect is replicated in a variety of studies, using different experimental paradigms and languages (see Petrova et al., 2011, Table 1, for an overview, but also Beyermann and Penke, 2014; Qu and Damian, 2016; Chen et al., 2016, for recent studies). Furthermore, Ziegler et al. (2003) demonstrate that not only the phonological but also the orthographic neighborhood size of a word has an impact on SWR. They report two opposing effects, the *inhibitory phonological effect*, and the *facilitatory orthographic effect*. Depending on a large phonological or orthographic neighborhood of a word, the SWR process is either impeded or facilitated.

There is still a debate on how orthography exactly influences the process of SWR. However, there are two prominent hypotheses about the source of orthographic effects in SWR (Pattamadilok et al., 2014). According to the *online hypothesis*, orthographic representations are co-activated *during* phonological processing, whereas the *offline hypothesis* claims that phonological representations change through the acquisition of reading and writing such that they also incorporate orthographic information.

In what follows, we present two models of SWR using a long short-term memory (LSTM) architecture (Hochreiter and Schmidhuber, 1997) and distributed representations, while focusing on German as a language. Our major outcomes are: (1) We design two models of SWR that instantiate the *offline* and the *online hypothesis* on the source of orthographic effects, respectively. (2) We replicate the *inhibitory phonological* and *facilitatory orthographic effect*, showing that these models are able to reproduce human-like behavior. (3) We provide testable hypotheses for future research based on the models' behavior, which allows us to further validate the *online* or *offline hypothesis*.

2 Methodology

2.1 Model architectures

We propose a recurrent model of SWR that consists of an LSTM that takes a sequence of phonemes as input and produces a meaning representation as output. The procedure of processing, e.g., the German word *Maus* (*mouse*) is illustrated in Figure 1. First, the model takes the respective phonemic sequence of [/m/, /aʊ/, /s/] as input. Then, it should build a vector representation that corresponds to a phoneme sequence, thus the phonological form of the entire word, to then produce a word meaning representation as output. This meaning representation should be as close as possible to the actual ground truth, which is the word embedding of *Maus* (*mouse*).

Phoneme embeddings learn the phonemic distribution well and implicitly capture articulatory distinctive features of phonemes (Silfverberg et al., 2018; Kolachina and Magyar, 2019). Therefore, phoneme vector representations are trained using word2vec (Mikolov et al., 2013) on the phonetic
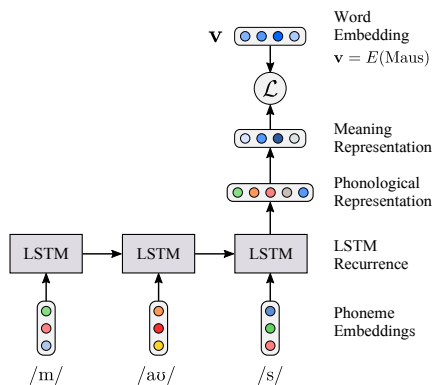

Figure 1: Sketch of a recurrent neural model of SWR.

transcription of the NEGRA corpus (Skut et al., 1997). The transcription is generated with the grapheme-to-phoneme converter tool provided by the Bavarian Archive for Speech Signals (BAS) (Reichel, 2012, 2014). The cbow model and negative sampling is used with window size 1 to obtain 30-dimensional phoneme embeddings.

Word meanings are approximated by word embeddings. We use pre-trained German fastText embeddings (Grave et al., 2018) as the output meaning representations of our models (see also Baayen et al., 2019; Chuang et al., 2020; Hendrix and Sun, 2020, for the similar use of word embeddings as semantic representations in models of word recognition).

The offline model The first architecture implements the theoretical assumption that a prelexical phonemic representation is mapped onto a lexical meaning representation, without incorporating explicit orthographic representations at the prelexical level. The offline model, which instantiates the *offline hypothesis*, processes one phoneme per time step. After the last phoneme of a phonological sequence is processed, a linear transformation is performed on the output of the LSTM layer which consists of 400 units. The resulting fully connected layer has 400 neurons and is then connected to the output layer. A tangent activation function is used on the output layer (300 units).

The online model The second proposed model architecture includes explicit orthographic information at the prelexical level, instantiating the *online hypothesis*. The online model processes two kinds of inputs – a sequence of 30-dimensional phoneme representations and a localist orthographic representation of a word that is based on character bigrams (818 units). The first input layer (30 units) is connected to an LSTM cell (400 units) which is fully connected to an intermediate layer (400 units). This intermediate layer is connected to an intermediate phonological layer (400 units). A tangent non-linearity is then used on it. On the other side of the model, a linear transformation together with a tangent non-linearity is applied on the second input layer to obtain a 100-dimensional layer. The intermediate phonological and orthographic representation are concatenated to a 500-dimensional vector which is then fully connected to a hidden layer of size 300. This hidden layer serves as an intermediate processing stage that processes both

types of information, auditory and visual ones, to then give the 300-dimensional meaning representation as output.

2.2 Training

A good model should be able to learn the meaning of spoken words seen during training and generalize to similar but unseen words. We expect the model to learn that very similar sounding words have a very similar meaning (e.g., *duck* and *ducks* share nearly the same semantic concept of a water bird with short legs). By training the model on inflected forms and lemmas, e.g. *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)*, even if it never encountered the phonological sequence and word meaning representation during the training phase.

For the training and test data, the most frequent singular and plural nouns in nominative case are extracted from the German Morphology Lexicon (Lezius, 2000), leading to 3118 inflected forms and their lemmas, as well as 583 single inflected forms in the training set, and their corresponding 583 testing lemmas in the test set. In this data set, a lemma is always one of the ten nearest neighbors (measured by cosine similarity) of its inflected form such that the meaning representations of an inflected form and the respective lemma are similar to each other in the embedding space.

The offline model is trained for 100 and the online model for 150 epochs, using the Adam optimizer with its default parameters in PyTorch, as well as the CosineEmbeddingLoss to minimize the cosine distance between the output of a model and the correct word embedding.

2.3 Evaluation

To evaluate the models, the cosine similarity between a model's output and every possible ground truth vector representation is computed. The set of competing word vectors, therefore, consists of 3701 word embeddings during training, and of 4284 (3701 training + 583 testing) vectors during testing. Given these competing word embeddings, Recall@k (R@k) is computed as the proportion of times that the set of top k 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, then this contributes to R@1. Furthermore, a word contributes to R@5 (R@10), if the corresponding ground truth word embedding is within the top 5 (top 10) most similar words to the output vector.

2.4 Simulation data

The model is considered to be successful if it can reproduce human behavioral data that is measured by Ziegler et al. (2003) in an auditory lexical decision task. The stimuli either have a large (+) or a small (-) number of phonological (PN) and orthographic neighbors (ON), which leads to the four categories ON-PN-, ON+PN-, ON-PN+, and ON+PN+. A word is considered to be an orthographic (phonological) neighbor of a target item if it is possible to create it by substituting one letter (one phoneme) in the target word (Coltheart's N, Coltheart et al., 1977). For example, *tape* is an orthographic neighbor of *type*, whereas /paɪp/ (*pipe*) is a phonological neighbor of /taɪp/ (*type*). The authors report two different effects on SWR.

The inhibitory phonological effect A large phonological neighborhood size impedes accessing the correct meaning representation of a word; whenever a stimulus has a large phonological neighborhood size (PN+), the reaction time in a downstream task like lexical decision is larger compared to a word that has a small phonological neighborhood size (PN-). A model should thus also have more difficulties to get to the correct word meaning representation for PN+ vs. PN- words.

The facilitatory orthographic effect Words with a large orthographic neighborhood size (ON+) produce shorter reaction times than words with a small orthographic neighborhood size (ON-). A large orthographic neighborhood size, therefore, facilitates SWR. Therefore, it should be easier for a model to produce the correct meaning representation for an ON+ compared to an ON- word.

2.5 Linking hypothesis

In a lexical decision task, shorter reaction times are associated with fast and effortless processing which is a result of strong word activations (Scharenborg and Boves, 2010). As word activation is assumed to be dependent on the degree of match between processed and stored information in the SWR process (Weber and Scharenborg, 2012), we infer the response time by comparing the model's output (processed information) with the ground truth representation of a word (stored information). A large difference would, therefore, indicate a relatively

weak word activation, which suggests a larger response time. On the other hand, a smaller error signals a stronger word activation, which corresponds to a smaller reaction time.

A larger error score for PN+ vs. PN- words, thus, corresponds to the *inhibitory phonological effect*, as a large phonological neighborhood size (PN+) impedes accessing the correct meaning representation of a word. By contrast, a large orthographic neighborhood size (ON+) facilitates the word recognition process. Hence, a lower error score for ON+ vs. ON- words is assumed to be an analog for the *facilitatory orthographic effect*.

3 Experiments

3.1 Word meaning retrieval task

After training, the models are evaluated on the training and the test set to compute the training and testing recall (Table 1). Training recall is nearly perfect for both models, showing that they are able to memorize the data well. However, the online model achieves a higher R@1 of 100% than the offline model in the training data. Overall, both models perform well in the word meaning retrieval task, which concerns activating the correct meaning representation based on a phonological word form.

3.2 Generalization task

On the test set, the offline model reaches an R@10 of 62.95%, an R@5 of 56.78%, and an R@1 of 21.61%, whereas the online model again performs comparatively better with a testing recall of 70.67% for R@10, 59.35% for R@5, and 22.98% for R@1. This is very good, given that the models have never encountered the exact phonological sequence, nor the word embedding of a testing item during training. The generalization performance of the models is an indicator that they globally learn how word forms and their semantics relate to each other. As for future work, one can compare these results with the performance of the models on unseen words which are semantically unrelated to those in the training set. Considering both training and testing recall values, the online model performs comparatively better in learning the meaning of spoken words. However, it still needs to be verified to what extent each of the models is able to reproduce human-like behavior.

Model	Split	R@10	R@5	R@1
Offline	Train	100	100	99.32
	Test	62.95	56.78	21.61
Online	Train	100	100	100
	Test	70.67	59.35	22.98

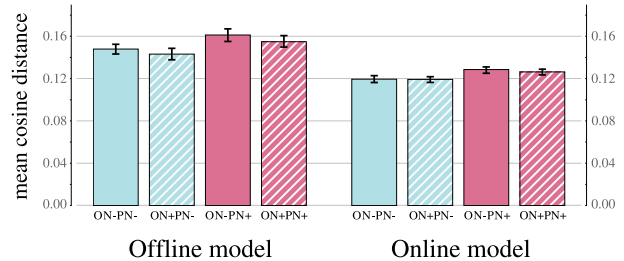Table 1: Training and testing recall in percent.

Figure 2: Mean cosine distance between the outputs and ground truths of the items of the four neighborhood categories. Error bars show standard errors.

3.3 Simulation task

To simulate the study by Ziegler et al. (2003), their experimental design is mimicked by dividing the German training data into the four neighborhood categories ON-PN-, ON+PN-, ON-PN+, and ON+PN+. Analogous to their categorisation, a word is considered to be part of the ON- category, when it has zero or one orthographic neighbor, otherwise it belongs to ON+. If a word has less than 3 phonological neighbors, it belongs to the PN- category, otherwise, it is considered to be part of the PN+ condition. For each of these four groups, we sample 70 items with similar mean word length, frequency, and density of the embedding space. The frequency of a word is estimated using the module *wordfreq* (Speer et al., 2018), whereas the density of the semantic space is approximated by subtracting the cosine distance between the ground truth word embedding and the mean vector of its ten nearest neighbors from 1.

Figure 2 shows a bar plot for each model that presents the mean cosine distance between the model's output of each word and the corresponding ground truth per condition after the models have been trained. For both models, the mean cosine distance is higher in the conditions with a large phonological neighborhood size (ON-PN+ and ON+PN+, pink bars in Figure 2) compared to the conditions with a low phonological neighborhood size (ON-PN- and ON+PN-, turquoise bars in Figure 2). This corresponds to a relatively lower word activation for PN+ items, indicating higher reaction times.

Thus, both models can reproduce the *inhibitory phonological effect*. A large orthographic neighborhood size (ON+PN- and ON+PN+, striped bars in Figure 2) has a beneficial impact on the models' performance. The mean cosine distance within the ON+PN- condition is lower compared to the ON-PN- group and it is also lower for the ON+PN+ compared to the ON-PN+ condition. This corresponds to the *facilitatory orthographic effect* and can also be observed for both model architectures. It is larger in the offline model which is surprising, because as opposed to the online model, it has no access to orthographic information. As the offline model instantiates the *offline hypothesis* which claims the phonological representation themselves contain implicit orthographic information, it is investigated whether also the phonological sequences of the training items reveal information about orthography which could have a beneficial effect on a model's performance.

Analysis of orthographic information A *friend* of a target word is a word that has the same rhyme and the same rhyme spelling, whereas *enemies* are words that have the same rhyme, but a different rhyme spelling (Ziegler et al., 2004). Therefore, words that have friends but zero enemies naturally fall into the category of consistent words (see Section 1), whereas words that have at least one enemy can be considered as being inconsistent. Based on the phonological sequence of a consistent word, one can infer its orthographic form, as its rhyme is always spelled in only one way. Therefore, consistent words provide implicit orthographic information in their phonological forms. An analysis of the friends and enemies in the training data reveals that the majority of items in the two groups with a large orthographic neighborhood, ON+PN- and ON+PN+, are consistent words. Furthermore, the mean error score for all consistent (253) and inconsistent words (62) in the training data (see Figure 3), shows that it is easier for the offline model to produce a good lexical meaning representation whenever a word is consistent, compared to inconsistent words that do not reveal reliable orthographic information. By contrast, the online model is not influenced by consistency. Therefore, the underlying reason for the *facilitatory orthographic effect* in the offline model is likely to be the phonology-orthography-consistency, rather than the size of the orthographic neighborhood.
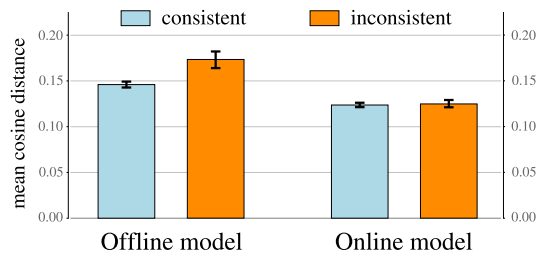
To assess whether consistency is an explanatory

Figure 3: Mean cosine distance between the outputs and ground truths of consistent and inconsistent words. Error bars show standard errors.
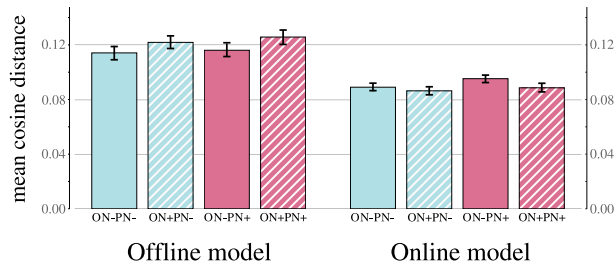

Figure 4: Mean cosine distance between the outputs and ground truths of Finnish items. Error bars show standard errors.

factor for the *facilitatory orthographic effect*, we eliminate the difference between consistent and inconsistent words by training the models on Finnish data. Finnish has a grapheme to phoneme mapping that is nearly one to one which leads to little to no inconsistent words (Joshi and Aaron, 2016).

Excluding the factor of consistency For the Finnish training data, the 2378 most frequent words are extracted from the vocabulary of the Finnish fastText embeddings (Grave et al., 2018). For the input of the models, Finnish phoneme embeddings are trained on the transcription of Finnish news texts (Newscrawl 2017, Goldhahn et al., 2012). Finnish fastText embeddings are used as meaning representations, as well as 540-dimensional localist orthographic representations within the online model. Four balanced samples of size 70 that correspond to the four neighborhood groups are drawn from the training data to then monitor the mean error score of each model per condition (see Figure 4).

The results after training the offline model on Finnish data show an inverse pattern compared to the German results. The offline model would, therefore, predict that no *facilitatory orthographic effect* can be observed in a lexical decision task with Finnish participants as every phonological sequence is nearly equally informative w.r.t. or-

thographic information. If this prediction proves true, this would further validate the *offline hypothesis* on the source of orthographic effects. For the online model, the general order of error scores is similar across languages. As it is not affected by consistency, the online model can also reproduce the *facilitatory orthographic effect* in Finnish. If this effect can be observed in a lexical decision task with Finnish participants, this would further validate the online model as a plausible model SWR, as well as the *online hypothesis*.

4 Conclusion

In this work, we propose two models of SWR that instantiate either the *online* or the *offline hypothesis* on the source of orthographic effects. We show that both models perform well in word meaning retrieval and in simulating the *inhibitory phonological* and *facilitatory orthographic effect*. The online model achieves the best training and testing performance, and shows the same pattern of results independent of the language of the data. It is not influenced by consistency, which indicates that the size of the orthographic neighborhood is at the origin of the *facilitatory orthographic effect* under the *online hypothesis*. This contrasts with the offline model that produces an orthographic consistency effect. When words don't differ in their consistency, the *facilitatory orthographic effect* is not present, which suggests that consistency is the underlying mechanism for this effect under the *offline hypothesis*. The models predict mutually exclusive outcomes in a lexical decision task in a language like Finnish that has a high phonology-orthography consistency. By testing these predictions, further evidence for either the *offline* or the *online hypothesis* can be provided.

Acknowledgments

References

Rolf Harald Baayen, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019.

Sandra Beyermann and Martina Penke. 2014. The Impact of Orthographic Consistency on German Spoken Word Identification. *International Journal of Disability, Development and Education*, 61(3):212–224.

Wei-Fan Chen, Pei-Chun Chao, Ya-Ning Chang, Chun-Hsien Hsu, and Chia-Ying Lee. 2016. Effects of orthographic consistency and homophone density on Chinese spoken word recognition. *Brain and Language*, 157-158:51–62.

Yu-Ying Chuang, Marie Lenka Vollmer, Elnaz Shafaei-Bajestan, Susanne Gahl, Peter Hendrix, and Rolf Harald Baayen. 2020. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior research methods*, pages 1–32.

Max Coltheart, Eileen Davelaar, Jon Torfi Jonasson, and Derek Besner. 1977. Access to the internal lexicon. In *Attention and performance IV*, pages 535–555. Hillsdale, NJ: Erlbaum.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 759–765, Istanbul, Turkey. European Languages Resources Association (ELRA).

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Peter Hendrix and Ching Chu Sun. 2020. A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

R. Malatesha Joshi and P. G. Aaron. 2016. *Handbook of Orthography and Literacy*. Routledge.

Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about Phonology? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 160–169, Florence, Italy. Association for Computational Linguistics.

Wolfgang Lezius. 2000. Morphy-german morphology, part-of-speech tagging and applications. In *Proceedings of the 9th EURALEX International Congress*, pages 619–623. University of Stuttgart.

William Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25:71–102.

William Marslen-Wilson and Lorraine Komisarjevsky Tyler. 1980. The temporal structure of spoken language understanding. *Cognition*, 8(1):1 – 71.

William Marslen-Wilson and Alan Welsh. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10:29–63.

James L McClelland and Jeffrey L Elman. 1986. The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1 – 86.

Tomas Mikolov, Kai Chen, G. S. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR 2013*.

Dennis Norris. 1994. Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52(3):189 – 234.

Chotiga Pattamadilok, José Morais, Cécile Colin, and Regine Kolinsky. 2014. Unattentive speech processing is influenced by orthographic knowledge: Evidence from mismatch negativity. *Brain and Language*, 137:103–111.

Ana Petrova, Gareth Gaskell, and Ludovic Ferrand. 2011. Orthographic Consistency and Word-Frequency Effects in Auditory Word Recognition: New Evidence from Lexical Decision and Rime Detection. *Frontiers in psychology*, 2:263.

Qingqing Qu and Markus Damian. 2016. Orthographic effects in spoken word recognition: Evidence from Chinese. *Psychonomic Bulletin & Review*, 24.

Uwe D. Reichel. 2012. PermA and Balloon: Tools for string alignment and text processing. In *Proc. Interspeech*, page 4 pages, Portland, Oregon.

Uwe D. Reichel. 2014. Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In R. Hoffmann, editor, *Elektronische Sprachverarbeitung 2014*, volume 71, pages 42–49. TUDpress, Dresden, Germany.

Odette Scharenborg and Lou Boves. 2010. Computational modelling of spoken-word recognition processes: Design choices and evaluation. *Pragmatics and Cognition*, 18:136–164.

Miikka P. Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. Sound Analogies with Phoneme Embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In *Fifth Conference on Applied Natural Language Processing*, pages 88–95, Washington, DC, USA. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

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

Johannes Ziegler and Ludovic Ferrand. 1998. Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, 5:683–689.

Johannes Ziegler, Ludovic Ferrand, and Marie Montant. 2004. Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks. *Memory & cognition*, 32:732–41.

Johannes Ziegler, Muneaux Mathilde, and Jonathan Grainger. 2003. Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, 48:779–793.