Modeling Languages with Their Own Parameters: A Response to subs2vec

First Author[1] & Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. First Author: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: my@email.com

Abstract

subs2vec (van Paridon & Thompson, 2021) provides word embeddings for 55 languages, derived from the Open Subtitles (Lison & Tiedemann, 2016) and Wikipedia (Wikimedia Downloads, 2018) corpora. However, these models were generated using the same computational parameters for all languages, without adjusting key hyperparameters such as minimum word frequency, vector dimension, or context window size. Prior work (Mandera et al., 2017) indicates that optimal parameters can differ across languages—for example, English and Dutch perform best at different dimensions and window sizes. In this study, we replicate the general approach of van Paridon and Thompson, but optimize embeddings for each language individually using the same corpora. Model quality is evaluated using published lexical norms (e.g., age of acquisition, valence, imageability, concreteness) as benchmarks, selecting the best-performing configuration per language. We present the results, examine the assumption of cross-linguistic similarity in embedding structure, and release all embeddings, code, and tools as an open package for researchers.

Keywords: embeddings, psycholinguistics, modeling

Modeling Languages with Their Own Parameters: A Response to subs2vec

The scientific study of language, or linguistics, has long sought to uncover the mechanisms and principles underlying human communication. From the early descriptive approaches of Boas (2013), first published in 1911, to the generative frameworks introduced in 1957 by Chomsky (Chomsky, 2002), linguistic theory has aimed to define the structure and function of natural languages. The evolution of the field has paralleled broader developments in cognitive science, computational modeling, and neuropsychology, establishing language as a central topic for interdisciplinary research Wilks (2006). In contemporary linguistics, a prominent area of focus lies in the computational modeling of language using large-scale corpora and machine learning techniques. Early efforts focused on machine translation and text analysis (K. S. Jones, 1994), while subsequent developments addressed tasks such as word-sense disambiguation, syntactic parsing, and sentiment analysis Medhat, Hassan, & Korashy (2014). Further computational approaches leverage algorithms to create numerical representations of words, phrases, and sentences, known as word embeddings. Models such as Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and fastText (Bojanowski, Grave, Joulin, & Mikolov, 2016) exemplify the integration of statistical methods into linguistic research, transforming the study of lexical semantics, syntactic structure, and discourse analysis.

Linguistic data, fundamental to natural language processing research, encompasses diverse forms ranging from raw, unprocessed text to human-provided subjective ratings and computationally enhanced attributes. This data underpins a wide array of tasks, including statistical analyses of word usage, such as lexical diversity measures (Bird, Klein, & Loper, 2009) and readability prediction (Pitler & Nenkova, 2008), as well as advanced applications like text generation (Clark, Ji, & Smith, 2018) and machine translation (Koehn, 2005). Moreover, linguistic datasets drive experiments across disciplines, supporting research in neurophysiology (Pereira et al., 2018), sociology (Garg, Schiebinger, Jurafsky, & Zou, 2018), and psychology (Paridon & Thompson, 2021). The subsequent section will delve into three critical categories of linguistic data: corpora, which provide structured collections of text; objective norms, which quantify measurable linguistic attributes; and subjective norms, which capture human perceptions and evaluations of language.

Corpora

Corpora, structured collections of text, serve as fundamental resources for linguistic and computational research, enabling systematic analysis of language data (Johansson & Oksefjell, 1998). A corpus typically consists of tokens—unique instances of word types—that are arranged within a principled format to facilitate linguistic studies (Ogden, Richards, & Malinowski, 2013). These collections may contain raw text, metadata, or annotated linguistic features, providing valuable insights into language usage, syntax, and semantics (Bird et al., 2009). Notable examples include Project Gutenberg, which offers a vast library of public domain texts for statistical analyses, and curated resources like the Brown Corpus, which categorize prose into diverse linguistic domains to support tasks such as frequency estimation and part-of-speech tagging Gerlach & Font-Clos (2020).

Wikipedia, an open-source, community-maintained encyclopedia, has emerged as one of the most extensively used corpora in linguistic research. With over six million articles in English and substantial coverage in other languages, Wikipedia supports a wide range of applications, including information retrieval, ontology development, and natural language processing tasks Medelyan, Milne, Legg, & Witten (2009). Its breadth and structured format make it an invaluable resource for creating large-scale language models and analyzing lexical semantics Mandera, Keuleers, & Brysbaert (2017). Wikipedia data is refreshed regularly and distributed as compressed XML files, ensuring reproducibility and access to current knowledge repositories.

The OpenSubtitles corpus, comprising over three million subtitles from films and television episodes in more than 60 languages, provides a rich source of pseudo-conversational linguistic data (Lison & Tiedemann, 2016). Subtitles are particularly valuable for studying spoken-like language, offering insights into lexical frequency, contextual usage, and semantic nuances (Brysbaert & New, 2009). The corpus is periodically updated and distributed in XML format, making it accessible for diverse research applications, including lexical complexity analysis and neural dialog generation Nakamura, Sudoh, Yoshino, & Nakamura (n.d.). As a multilingual resource, OpenSubtitles has been instrumental in advancing computational models for less-studied languages and cross-linguistic analyses.

Objective Data

Objective lexical norms capture measurable features of words, such as word length, syllable count, and phonological or orthographic neighborhoods, which are groups of words sharing similar linguistic attributes Marian (2017). These norms are critical for exploring language structure, semantic memory, and bilingual lexical storage (E. M. Buchanan, Valentine, & Maxwell, 2019). Tools like SUBTLEX-UK calculate frequency-based metrics, demonstrating, for example, that higher-frequency verb conjugations are processed faster than irregular forms Bowden, Gelfand, Sanz, & Ullman (2010). While frequency provides a core measure of lexical accessibility, other corpus-derived metrics capture the breadth and variability of word use. Contextual diversity, the number of distinct contexts in which a word appears, often predicts lexical processing as well as or better than frequency (Adelman, Brown, & Quesada, 2006). Semantic diversity indexes the variability of a word's usage across contexts and is linked to effects of ambiguity and polysemy (Hoffman, Lambon Ralph, & Rogers, 2013). Finally, word length remains a robust factor that interacts with frequency and neighborhood structure (Brysbaert et al., 2011).

Phonological and orthographic neighborhoods, which respectively include similar-sounding and visually similar words, play a role in word recognition and production L. Buchanan, Westbury, & Burgess (2001). For instance, words in dense phonological neighborhoods are recognized and produced more efficiently Taler, Aaron, Steinmetz, & Pisoni (2010). Likewise, semantic neighborhood density reflects how many words in a semantic space have meanings similar to a given word. Using distributional models such as BEAGLE (M. N. Jones & Mewhort, 2007), researchers can estimate how crowded a word's "meaning neighborhood" is, providing a meaning-level counterpart to phonological and orthographic neighborhoods and capturing competition or facilitation effects based on semantic similarity (M. N. Jones, Johns, & Recchia, 2012). Normative measurements like word frequency, lexical diversity, and sentence complexity inform linguistic richness and proficiency (Malvern, Richards, Chipere, & Durán, 2004). Combined with metrics such as the Flesch-Kincaid Readability Test, these norms provide valuable insights into vocabulary development and document complexity (Flesch, 1948). These objective norms, therefore, serve as foundational tools for studying linguistic phenomena and assessing language proficiency.

Subjective Data

Subjective lexical norms are derived through human ratings and capture perceptual, emotional, and experiential attributes of words. These norms include age of acquisition, familiarity, imageability, concreteness, valence, and arousal, among others (E. M. Buchanan et al., 2019). For instance, age of acquisition measures when a word is typically learned and aids in predicting word recognition times Brysbaert & Ghyselinck (2006). Familiarity gauges how common a word is within an individual's experience and often correlates with frequency of exposure, influencing long-term priming effects (Ray & Bly, 2007). Similarly, imageability captures how easily a word evokes a mental image, significantly impacting word recognition and recall (Boukadi, Zouaidi, & Wilson, 2016). Concreteness reflects how closely a concept relates to a physical object, with concrete words eliciting faster responses in lexical decision tasks compared to abstract words Barber, Otten, Kousta, & Vigliocco (2013). Emotional dimensions, such as valence (pleasantness) and arousal (emotional intensity), are integral to affective priming tasks, where response times are influenced by the congruence of priming and target word valence Warriner, Kuperman, & Brysbaert (2013).

Databases containing subjective norms, such as the MRC Psycholinguistic Database and the Linguistic Inquiry and Word Count (LIWC) system, integrate both objective and subjective lexical ratings Tausczik & Pennebaker (2010). These resources enable researchers to study emotional, cognitive, and social aspects of language. For example, LIWC categorizes words into linguistic and emotional categories, such as "anger" or "sadness," based on iterative human review (Tausczik & Pennebaker, 2010). Such databases are pivotal for psycholinguistic and computational studies, as they provide standardized measures for analyzing the interplay of lexical properties and human perception. By combining objective measures like frequency and subjective dimensions like valence, these tools offer comprehensive insights into language processing and its cognitive underpinnings.

Linguistic Modeling

Computational modeling of linguistic data has evolved significantly over the decades, beginning with early approaches such as Latent Semantic Analysis (LSA) in the 1990s. LSA represented words and contexts in a high-dimensional space derived from a co-occurrence matrix, using techniques like Singular Value Decomposition to reduce dimensionality and emphasize meaningful relationships between words (Landauer & Dumais, 1997). These foundational methods introduced the concept of vectorizing language for analysis, enabling researchers to explore semantic relationships through spatial proximity in vector space Sahlgren (2006). However, these early models, often called "bag-of-words" approaches, treated words as discrete entities, overlooking word order and internal word structures, which limited their ability to capture nuanced linguistic patterns (Mikolov et al., 2013).

The introduction of neural network-based methods in the 2010s marked a turning point in computational linguistics. Mikolov et al. (2013) developed word2vec, which utilized two novel algorithms—Skip-Gram (SG) and Continuous Bag of Words (CBOW)—to predict word context and improve upon earlier models' efficiency and scalability. These innovations allowed for the creation of embeddings from datasets containing billions of words, with enhanced representation in higher-dimensional spaces. Building on this foundation, Bojanowski et al. (2016) introduced fastText, incorporating subword information to represent internal word structures, enabling the handling of out-of-vocabulary tokens. The development of these models, along with frameworks like *gensim* (Řehůřek & Sojka, 2010), consolidated disparate techniques into accessible software packages, making computational modeling of language more efficient and widely applicable. These advancements have paved the way for analyzing large-scale corpora and predicting complex linguistic and cognitive norms, revolutionizing natural language processing and related fields.

subs2vec

van Paridon and Thompson (2021) developed word embedding models derived from spoken language across multiple languages, utilizing the OpenSubtitles corpus (Lison & Tiedemann, 2016) and the fastText implementation of word2vec. Their work emphasized the importance of spoken language corpora, which better approximate language acquisition and usage compared to written text, addressing a limitation of prior studies that predominantly relied on Wikipedia-based corpora (Al-Rfou, Perozzi, & Skiena, n.d.). Models of combined resources were found to predict subjective norm ratings across multiple languages, such as concreteness, valence, and arousal, suggesting that complementary resources are useful for modeling linguistic data.

The models developed by van Paridon and Thompson were constructed using data from 55 languages with uniform parameters across all corpora, regardless of size or linguistic structure. While this consistency aids in cross-linguistic comparisons, other research suggests that model performance can vary significantly based on parameter optimization. For instance, Mandera et al. (2017) demonstrated that the choice of parameters, such as vector dimensionality and window size, affects the quality of word embeddings, with optimal settings differing between languages. Their findings highlight that English embeddings performed best with 300 dimensions and a window size of six, whereas Dutch embeddings achieved superior results with 200 dimensions and a window size of ten. These results challenge the assumption that a uniform parameter set is equally effective across languages, given the structural and typological diversity of linguistic systems. This research considers the necessity of tailoring model parameters to individual language characteristics and research goals to enhance the accuracy and applicability of multilingual word embeddings.

The Current Study

To examine the implicit assumption that all languages can be effectively represented using identical word embedding model parameters, we will construct matrices across a range of parameter combinations, including vector dimensions (50, 100, 200, 300, 500), window sizes (1, 2, 3, 4, 5, 6), and embedding algorithms (Continuous Bag of Words [CBOW] and Skip-Gram). The selected dimensional values reflect those commonly utilized in linguistic studies, as highlighted by Mandera et al. (2017). Window sizes were constrained to a maximum of six based on preliminary experimentation, which indicated that larger window sizes yielded negligible differences in predictive performance.

Unlike prior studies that imposed limitations on corpus size, such as Al-Rfou et al. (n.d.), who restricted corpora to 10,000 words, and van Paridon and Thompson (2021), who used corpora capped at 1 million words, our models will not limit corpus size. We will evaluate these models by testing their ability to predict:

1) a direct replication of the same norms used in van Paridon and Thompson,
2) objective normed data via word frequencies available for all languages
3) extension to subjective normed data available in more languages than present in the previous investigation

This approach will identify the optimal combination of parameters for each language, providing insights into how embedding models should be tailored for future cross-linguistic studies.

Method

Technical Implementation

The fastText model (Bojanowski et al., 2016) from the *genism* version 3.8.3 Python package (Řehůřek & Sojka, 2010) was used to generate the embeddings from the concatenated corpus files (described below). We varied the dimension (50, 100, 200, 300, 500), window size (1, 2, 3, 4, 5, 6), and algorithm parameters to the model (SG: SkipGram, CBOW: Continuous Bag of Words), while holding the remaining parameters constant. The dimensions, window size, and algorithm were chosen as the parameters of interest based on previous research showing they varied between datasets (Bojanowski et al., 2016; Mandera et al., 2017; Mikolov et al., 2013).

These parameter variations resulted in 60 possible combinations per language. Remaining parameter settings were matched to those used in the subs2vec experiment: 1) minimum word count: 5, 2) minimum length of subword ngram: 3, 3) maximum length of subword ngram: 6, 4) sampling threshold: .0001, 5) learning rate: .05, 6) rate of updating the learning rate: 100, 7) epochs: 10, 8) number of negatives sampled in the loss function: 10.

Figure 1 outlines the workflow for data acquisition, text preprocessing, corpus creation, and the generation of word-by-dimension matrices for each parameter combination. These procedures build on the original Python code from the subs2vec paper, with modifications tailored to the needs of this experiment. The full source code is available at https://github.com/SemanticPriming/word2manylanguages, and a working example of the pipeline can be found at XXCODE OCEAN HEREXX.

Data Acquisition. This experiment used datasets in 59 languages for which evaluation data were available. The language set includes those from the van Paridon and Thompson study, along with Japanese, Thai, Mandarin, and Cantonese. A full list of languages, along with unique sentence and token counts, is provided in Appendix A. Corpora were built from Wikipedia and OpenSubtitles archives. Open Subtitles files were downloaded from the URL http://opus.nlpl.eu/download.php?f=OpenSubtitles/v2018/raw/%7Blanguage%7D.zip, substituting the ISO3166 country code for {language}. The OpenSubtitles archive has updated since original download, but a working example of the file download is provided on the CODE OCEAN page. The OpenSubtitles files contain XML-formatted files for each movie or episode subdivided by year. The movie/episode names are not included in the data, and the order of the sentences is randomized to avoid copyright violation. Wikipedia is organized by language, with each language's content compiled into a single XML file containing article text and metadata. Wikipedia dump files were downloaded from http://dumps.wikimedia.your.org/%7Blanguage%7Dwiki/latest/%7Blanguage%7Dwiki-latest-pages-meta-current.xml.bz2, where {language} is the ISO 3166 code (e.g., en for English, de for German). The download dates for each archive are listed in Appendix A.

Data Processing. The downloaded data included markup in eXtensible Markup Language (XML), which was removed prior to corpus creation. Markup tokens do not reflect natural language content and can distort frequency counts (Bird et al., 2009). We used regular expressions to strip out markup elements such as tags, punctuation (parentheses, hyphens, apostrophes, slashes, etc.), links, and extraneous whitespace. For Wikipedia data specifically, additional elements like category labels, references, tables, and image tags were also removed. All text was lowercased to normalize the data.

van Paridon and Thompson (2021) applied sentence-level deduplication within each subtitle and Wikipedia document to reduce the influence of commonly repeated phrases. In contrast, we chose to retain these frequent phrases—such as "Thank you" because of their prevalence in spoken language, which we consider relevant to our analysis. We did apply document-level deduplication to avoid including exact duplicates, though given the curated nature of our data sources, the likelihood of such duplication was low. One corpus file was produced per language, with each file containing one sentence per line.

Data Analysis

The word embeddings generated during model training were evaluated based on their ability to predict psycholinguistic variables relevant to our research questions. For the direct replication (Research Question 1), we used the same norm datasets employed by Paridon and Thompson (2021). To enable evaluation across all languages, we also included word frequency prediction for Research Question 2 (see Brysbaert & New, 2009). Finally, we extended the analysis to additional normed datasets not used in the original study. These included a representative set of subjective norms: age of acquisition, valence, arousal, concreteness, and familiarity, selected for their widespread use in psycholinguistic research (Alario & Ferrand, 1999).

We used 10-fold ridge regression ($k = 10$) to predict norm values from the embeddings. Ridge regression was chosen due to its effectiveness in mitigating multicollinearity which is a common issue in word embedding models (Kaveh-Yazdy & Zarifzadeh, n.d.), and its demonstrated ability to improve mean squared error performance in this context (Yeh, Yeh, & Shen, 2020). This approach follows the evaluation procedure used in the original subs2vec study. The ridge regression alpha parameter, which controls the regularization strength, was set to the default value of 1 to balance bias and variance, consistent with prior work.

For each norm prediction task, we selected the simplest model whose R^2 value fell within 1% of the best-performing model. Simplicity was defined by the fewest embedding dimensions and the smallest window size. The adjusted R^2 value accounts for out-of-vocabulary coverage by multiplying the R^2 by the proportion of test words present in the embedding matrix. All model outputs are available in the supplemental materials. Given the volume of results, we developed a Shiny application (Chang et al., 2021) to help researchers explore and select optimal models for specific languages and variables of interest.

Results

Research Question 1

For the first set of evaluations, each language model was tested using the same normed datasets Paridon and Thompson (2021) (see Appendix B for the full list). We applied the same analysis approach as the original study which was ridge regression using the model's output vectors to predict norm values for matched tokens, as detailed in the data analysis section. Because all tables are very large, we recommend examining prediction for specific language, algorithm, and dataset combinations on our online resources or shiny application. The heatmaps shown in Figure 2 visualize the top three performing models for each algorithm. For CBOW, there is a clear trend favoring simpler models, with the most frequent configuration being 50 dimensions and a window size of 1. In contrast, the Skip-Gram results show a more balanced spread across dimensionalities, though still skewed toward smaller window sizes, most commonly size 3 or smaller. These results indicate that the parameter settings used in the original fastText models (300 dimensions, window size of 5) are unlikely to be optimal across languages. Notably, those original settings do not appear in the top three results for any language or prediction task under Research Question 1.

Research Question 2

The second set of tests addressed a key limitation: the lack of normed datasets for many of the languages modeled in this and previous studies. While prior work did not evaluate all available languages, word frequency was available for all models, and word frequency is known to correlate with numerous linguistic phenomena (Brysbaert & New, 2009). Unigram (i.e., single-token) frequency counts were directly extracted from the same corpora used to train the embeddings. However, this frequency data initially posed challenges: it included ligatures and diacritics that did not align with the normalized forms in the word-by-dimension matrices. To address this, we applied Unicode normalization and case folding, a standard approach for harmonizing case and character representation in internationalized text. Despite these steps, a substantial number of words remained unmatched, primarily due to the minimum frequency threshold of five tokens in the word embeddings, in contrast to no such threshold in the raw unigram frequency data.

To evaluate performance, frequency data from Wikipedia and OpenSubtitles were analyzed separately, using the combined models built for this study. Full results can be found online and on our interactive shiny application. Note that negative R^2 values indicate a penalty for the model failing to represent words found in the frequency data but missing from the vector space. Overall, unigram frequencies proved more difficult to predict from word embeddings than normed psycholinguistic variables. This pattern may reflect the known challenges of estimating lexical properties from decontextualized representations, where static embeddings carry far less explanatory power than context-aware models (Ethayarajh, n.d.). These limitations likely contributed to lower predictive performance, with substantial variation across languages.

Figures 3 and 4 display the top-performing models for predicting frequencies, separated by algorithm. As in the norm prediction tasks, simpler models again dominated. CBOW strongly favored the combination of 50 dimensions and a window size of 1. Skip-gram results were more varied but still leaned toward lower dimensionalities and small window sizes. Interestingly, differences in predictive performance between Wikipedia and OpenSubtitles suggest that dataset type matters. While Paridon and Thompson (2021) showed that combining formal (Wikipedia) and informal (OpenSubtitles) corpora can improve overall model performance, our findings suggest that matching the style of the model's training data to the test data may be even more effective if practical. These results raise the possibility that corpus-specific models may outperform general-purpose models for certain applications.

Research Question 3

The third set of evaluations used datasets from the Linguistic Annotated Bibliography (E. M. Buchanan et al., 2019), which contain normed psycholinguistic data similar to those used in the replication set, but cover a broader range of languages and norm types. A full overview of these datasets is provided in our online materials and shiny application. As with previous tests, the results show substantial variation across languages, reinforcing the conclusion that no single parameter configuration performs best across all contexts. The heatmaps in Figure 5, separated by algorithm, reflect patterns similar to those observed in the replication analysis. CBOW models again favored simpler configurations, though the most common cluster shifted to 100 dimensions (compared to 50 in the earlier CBOW results). Skip-gram models remained more distributed, with a consistent preference for smaller window sizes but somewhat higher dimensionalities overall.

Discussion

This experiment demonstrates that the structure and parameterization of word embedding models significantly impact their performance, and that these effects vary across languages and tasks. While Paridon and Thompson (2021) showed that combining formal and informal language sources (Wikipedia and OpenSubtitles) improves predictive accuracy over models trained on Wikipedia alone, our findings go further: even with the same training data, the optimal embedding parameters, such as vector dimensionality and window size, differ markedly across languages and tasks. This finding reinforces the importance of customizing model configurations rather than relying on default or pre-trained settings.

Despite the rise of transformer-based models, recent studies have shown that classic word embedding approaches such as fastText remain competitive and in some cases outperform deep learning models on specific tasks (Wang, Nulty, & Lillis, 2020). These classic models are more interpretable, computationally efficient, and resource-accessible, making them ideal for researchers without access to extensive compute infrastructure. To support the broader research community, we have made available the full set of models and evaluation results from this study, along with open-source code for training and evaluating embeddings across languages and tasks. These materials will be particularly valuable for researchers working with lower-resource languages, or those conducting multilingual studies where model retraining from scratch may be impractical.

Across all three sets of evaluations, norm prediction replication, frequency prediction, and extended norm prediction using the Linguistic Annotated Bibliography datasets, our results consistently show wide variation in optimal parameters. These results answer our first research question: no single configuration generalizes well across languages. This result is consistent with linguistic theory. Languages differ not only in script and morphology, but also in typological features such as word order, determiners, affixation, and compounding (Dryer, 2013). These differences shape token distributions and word co-occurrence patterns, which are central to embedding learning. Additionally, corpus size and the lexical diversity of the source text contribute to how embeddings are learned, especially in low-resource or morphologically rich languages (Vania & Lopez, n.d.).

Our shiny application allows researchers to review the optimal parameter settings for each language and task. Even within a single language, the best settings for predicting word frequency in formal (Wikipedia) versus informal (OpenSubtitles) data diverged. Likewise, different psycholinguistic variables, such as valence, age of acquisition, and imageability, were best predicted by models with different configurations. This finding supports the view that parameter tuning should be context-dependent, aligned with both the source of the input data and the nature of the variable to be predicted. For example, emotional valence may be more prevalent in informal speech, while age of acquisition norms may be better reflected in formal, education-linked text.

Given the widespread use of word embeddings across psycholinguistics, natural language processing, and cross-linguistic studies, this variability has important implications. Many studies rely on pre-trained embeddings (e.g., from fastText or BERT) assuming they are broadly applicable. Our findings suggest caution in this approach. Researchers should consider re-training or fine-tuning embeddings using representative data and tuning parameters for their specific application. More systematic evaluation across languages, tasks, and variable types is needed to better understand these dependencies.

Limitations

While this study extended prior work by building and testing word embedding models for 59 languages, we were limited by the availability of validated norm datasets. Norms are difficult to obtain for many languages, especially those with fewer computational and psycholinguistic resources. Language resources are frequently published in journals like *Behavior Research Methods* and *Language Resources and Evaluation* and should continue to evolved with increased computational power. New big team science initiatives, such as the ManyLanguages collaboration (*ManyLanguages*, n.d.), can improve the availability and diversity of languages published for research use.

Another key limitation is that our evaluation was task-specific. While we tested prediction of norm variables and frequency data, other important tasks (e.g., analogy solving, named entity recognition, semantic similarity judgments) may yield different optimal configurations. Thus, while our models provide strong baselines for norm-based prediction, further tuning may be required for other applications. Finally, we were unable to identify consistent patterns linking optimal parameters to geographic proximity or language families (Research Question 3). While this may reflect the complex, multidimensional nature of linguistic structure and usage, it also suggests the need for deeper investigation, potentially incorporating sociolinguistic and typological data to uncover more subtle patterns.

Future Work

One promising direction for future research is to explore variation within languages, particularly across dialects. While this study included multiple dialects of Chinese (e.g., Mandarin and Cantonese), only one variant was used for most other languages, such as English, Spanish, and Portuguese. Previous research has shown that dialectal variation can significantly affect lexical choice, syntax, and even semantic interpretation (Blodgett, Green, & O'Connor, n.d.; Joshi et al., 2025), suggesting that word embeddings trained on different dialects may vary in both structure and performance. Future studies should investigate how embedding performance differs across dialects, particularly in languages with widespread regional variation. Additionally, expanding coverage to underrepresented languages, especially those from Africa, South Asia, and the Pacific, remains a critical goal. While initiatives like Masakhane (Nekoto et al., n.d.) and the AI4D project (Mann & Hilbert, 2020) have made progress in this area, many languages still lack sufficient corpora or standardized benchmarks for evaluation. As more multilingual and open-access corpora become available, we can begin to build and test embeddings for these languages and assess whether the same variability in optimal parameters holds. Finally, understanding why certain parameter configurations work better for particular languages or tasks remains an open question. Factors such as morphological complexity (Cotterell et al., n.d.), word frequency distributions (Brysbaert et al., 2011; Brysbaert & New, 2009), and syntactic structure (Dryer, 2013) likely influence embedding learning and performance. Future work that integrates computational modeling with linguistic typology could provide insights into the underlying mechanics of embedding optimization, helping to develop more universal or adaptive modeling strategies.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Alario, F.-X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, *31*(3), 531–552. https://doi.org/10.3758/BF03200732

Al-Rfou, R., Perozzi, B., & Skiena, S. (n.d.). *Polyglot: Distributed word representations for multilingual NLP*. https://doi.org/10.48550/ARXIV.1307.1662

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53. https://doi.org/10.1016/j.bandl.2013.01.005

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. "O'Reilly Media, Inc.".

Blodgett, S. L., Green, L., & O'Connor, B. (n.d.). *Demographic dialectal variation in social media: A case study of african-american english*. https://doi.org/10.48550/arXiv.1608.08868

Boas, F. (Ed.). (2013). *Handbook of american indian languages* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139626545

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv Preprint arXiv:1607.04606*.

Boukadi, M., Zouaidi, C., & Wilson, M. A. (2016). Norms for name agreement, familiarity, subjective frequency, and imageability for 348 object names in Tunisian Arabic. *Behavior Research Methods*, *48*(2), 585–599. https://doi.org/10.3758/s13428-015-0602-3

Bowden, H. W., Gelfand, M. P., Sanz, C., & Ullman, M. T. (2010). Verbal Inflectional Morphology in L1 and L2 Spanish: A Frequency Effects Study Examining Storage Versus

Composition. *Language Learning*, *60*(1), 44–87. https://doi.org/10.1111/j.1467-9922.2009.00551.x

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology*, *58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, *13*(7-8), 992–1011. https://doi.org/10.1080/13506280544000165

Brysbaert, M., & New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). LAB: Linguistic annotated bibliography – a searchable portal for normed database information. *Behavior Research Methods*, *51*(4). https://doi.org/10.3758/s13428-018-1130-8

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*(3), 531–544. https://doi.org/10.3758/BF03196189

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., … R), R. C. T. (tar. implementation from. (2021). *Shiny: Web application framework for r.* Retrieved from https://CRAN.R-project.org/package=shiny

Chomsky, N. (2002). *Syntactic structures.* Mouton de Gruyter. https://doi.org/10.1515/9783110218329

Clark, E., Ji, Y., & Smith, N. A. (2018). *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* 2250–2260. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1204

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A. D., … Hulden, M. (n.d.). *The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection*. https://doi.org/10.48550/arXiv.1810.07125

Dryer, M. S. (2013). Order of subject, object and verb (v2020.4) [Data set]. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Zenodo. https://doi.org/10.5281/zenodo.13950591

Ethayarajh, K. (n.d.). *How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings*. https://doi.org/10.48550/arXiv.1909.00512

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16). https://doi.org/10.1073/pnas.1720347115

Gerlach, M., & Font-Clos, F. (2020). A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics. *Entropy*, *22*(1), 126. https://doi.org/10.3390/e22010126

Griffiths, S., Purver, M., & Wiggins, G. (2015, November 4). *From Phoneme to Morpheme: A Computational Model*. Universität Tübingen. https://doi.org/10.15496/publikation-8639

Heuven, W. J. B. van, Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

https://doi.org/10.1080/17470218.2013.850521

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Johansson, S., & Oksefjell, S. (1998). *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Rodopi.

Jones, K. S. (1994). *Natural Language Processing: A Historical Review* (A. Zampolli, N. Calzolari, & M. Palmer, Eds.). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-0-585-35958-8_1

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *66*(2), 115–124. https://doi.org/10.1037/a0026727

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., & Dippold, D. (2025). Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, *57*(6), 149:1149:37. https://doi.org/10.1145/3712060

Kaveh-Yazdy, F., & Zarifzadeh, S. (n.d.). *Measuring economic policy uncertainty using an unsupervised word embedding-based method*. https://doi.org/10.48550/arXiv.2105.04631

Koehn, P. (2005, September 13). *MTSummit 2005*. 7986. Phuket, Thailand. Retrieved from https://aclanthology.org/2005.mtsummit-papers.11/

Kucera, H., Francis, W. N., Carroll, J. B., & Twaddell, W. F. (1967). *Computational analysis of present day american english* (First Edition). Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lison, P., & Tiedemann, J. (2016). *Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.*

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical Diversity and Language Development.* London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230511804

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Mann, S., & Hilbert, M. (2020). *AI4D: Artificial Intelligence for Development.* Retrieved from https://escholarship.org/uc/item/2qv3h863

ManyLanguages. (n.d.). Retrieved from https://many-languages.com/

Marian, V. (2017). Orthographic and phonological neighborhood databases across multiple languages. *Written Language & Literacy*, *20*(1), 6–26. https://doi.org/10.1075/wll.20.1.02mar

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 716–754. https://doi.org/10.1016/j.ijhcs.2009.05.004

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of W ikipedia. *Journal of the Association for Information Science and Technology*, *66*(2), 219–245. https://doi.org/10.1002/asi.23172

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* 31113119.

Nakamura, R., Sudoh, K., Yoshino, K., & Nakamura, S. (n.d.). *Another diversity-promoting objective function for neural dialogue generation.* https://doi.org/10.48550/ARXIV.1811.08100

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., … Bashir, A. (n.d.). *Participatory research for low-resourced machine translation: A case study in african languages.* https://doi.org/10.48550/arXiv.2010.02353

Ogden, C. K., Richards, I. A., & Malinowski, B. (2013). *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism.* Martino Fine Books.

Paridon, J. van, & Thompson, B. (2021). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, *53*(2), 629–655. https://doi.org/10.3758/s13428-020-01406-3

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., … Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963. https://doi.org/10.1038/s41467-018-03068-4

Pitler, E., & Nenkova, A. (2008). *the Conference.* 186. Honolulu, Hawaii: Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613742

Ray, S., & Bly, B. M. (2007). Investigating Long-Term Semantic Priming of Middle- and Low-Familiarity Category Exemplars. *The Journal of General Psychology*, *134*(4), 453–466. https://doi.org/10.3200/GENP.134.4.453-466

Řehůřek, R., & Sojka, P. (2010). *Software Framework for Topic Modelling with Large Corpora.* University of Malta. Retrieved from https://repozitar.cz/publication/15725/cs/Software-Framework-for-Topic-Modelling-with-Large-Corpora/Rehurek-Sojka

Sahlgren, M. (2006). *The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-1037

Sampson, G. (1980). *Schools of linguistics: Competition and evolution.* London: HarperCollins Publishers Ltd.

Schütze, H. (1992). *Word space. 5.* Morgan-Kaufmann. Retrieved from https://proceedings.neurips.cc/paper/1992/hash/d86ea612dec96096c5e0fcc8dd42ab6d-Abstract.html

Schwanenflugel, P. J. (1991). *Chapter 2 Contextual Constraint and Lexical Processing.* Elsevier. https://doi.org/10.1016/S0166-4115(08)61528-9

Smolenska, G., Kolb, P., Tang, S., Bitinis, M., Hernández, H., & Asklöv, E. (2021). *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).* 632–639. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.semeval-1.81

Taler, V., Aaron, G. P., Steinmetz, L. G., & Pisoni, D. B. (2010). Lexical Neighborhood Density Effects on Spoken Word Recognition and Production in Healthy Aging. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *65B*(5), 551–560. https://doi.org/10.1093/geronb/gbq039

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 2454. https://doi.org/10.1177/0261927X09351676

Vania, C., & Lopez, A. (n.d.). *From characters to words to in between: Do we capture morphology?* https://doi.org/10.48550/arXiv.1704.08352

Vitevitch, M. S., & Luce, P. A. (2016). Phonological Neighborhood Effects in Spoken Word Perception and Production. *Annual Review of Linguistics*, *2*(1), 75–94. https://doi.org/10.1146/annurev-linguistics-030514-124832

Wang, C., Nulty, P., & Lillis, D. (2020, December 18). *A comparative study on word embeddings in deep learning for text classification*. 37–46. https://doi.org/10.1145/3443279.3443304

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Wilks, Y. (2006). *Computational linguistics: history* (K. Brown, Ed.). Oxford: Elsevier. https://doi.org/10.1016/B0-08-044854-2/00928-7

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10. https://doi.org/10.3758/BF03202594

Yeh, H.-Y., Yeh, Y.-C., & Shen, D.-B. (2020). Word Vector Models Approach to Text Regression of Financial Risk Prediction. *Symmetry*, *12*(1), 89. https://doi.org/10.3390/sym12010089

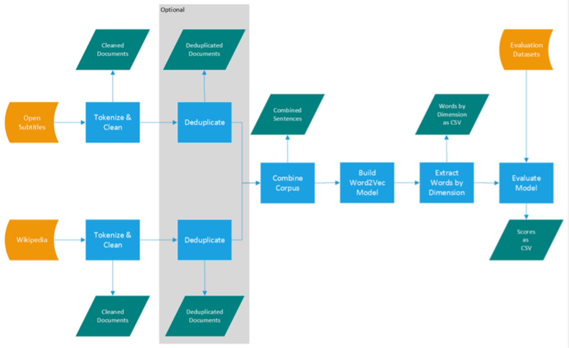

Figure 1. Flow chart representation of data processing, model creation, and prediction for this study.

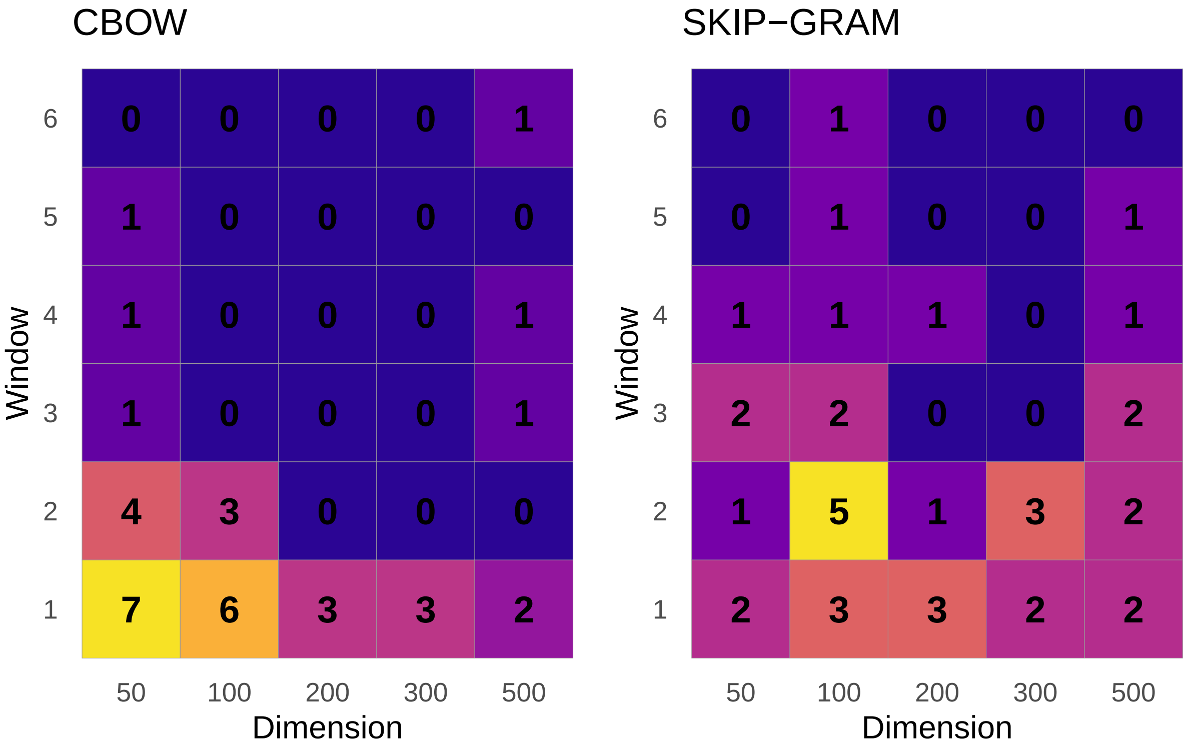

Figure 2. Heatmaps showing the number of languages achieving their best performance for each combination of embedding dimension (x-axis) and context window size (y-axis) for the CBOW (left) and skip-gram (right) algorithms. For CBOW, optimal configurations are concentrated at lower dimensions and smaller window sizes, whereas skip-gram tends to favor higher dimensions and moderately larger windows. Numbers inside tiles indicate the count of languages for that parameter combination.

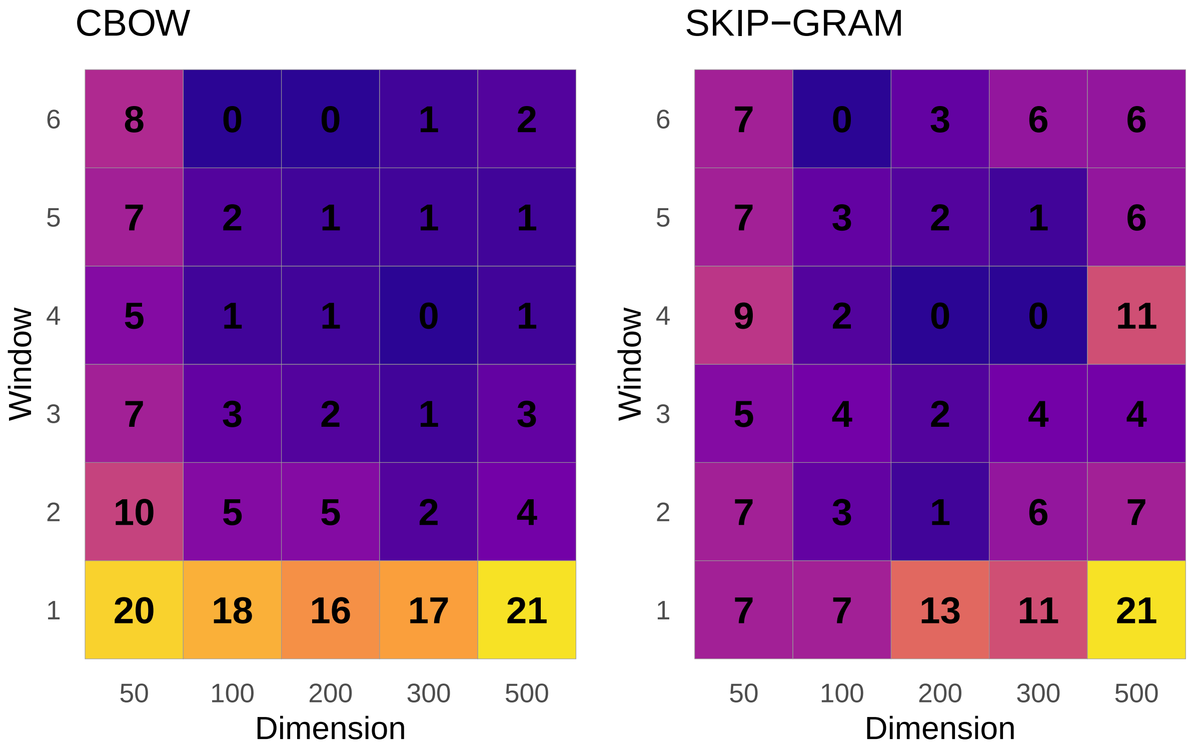

Figure 3. Heatmaps showing the number of languages for which each combination of embedding dimension (x-axis) and context window size (y-axis) yielded the best predictive performance for subtitle-based word frequency norms. Numbers within tiles indicate the count of languages achieving the top 3 highest prediction for that parameter combination.

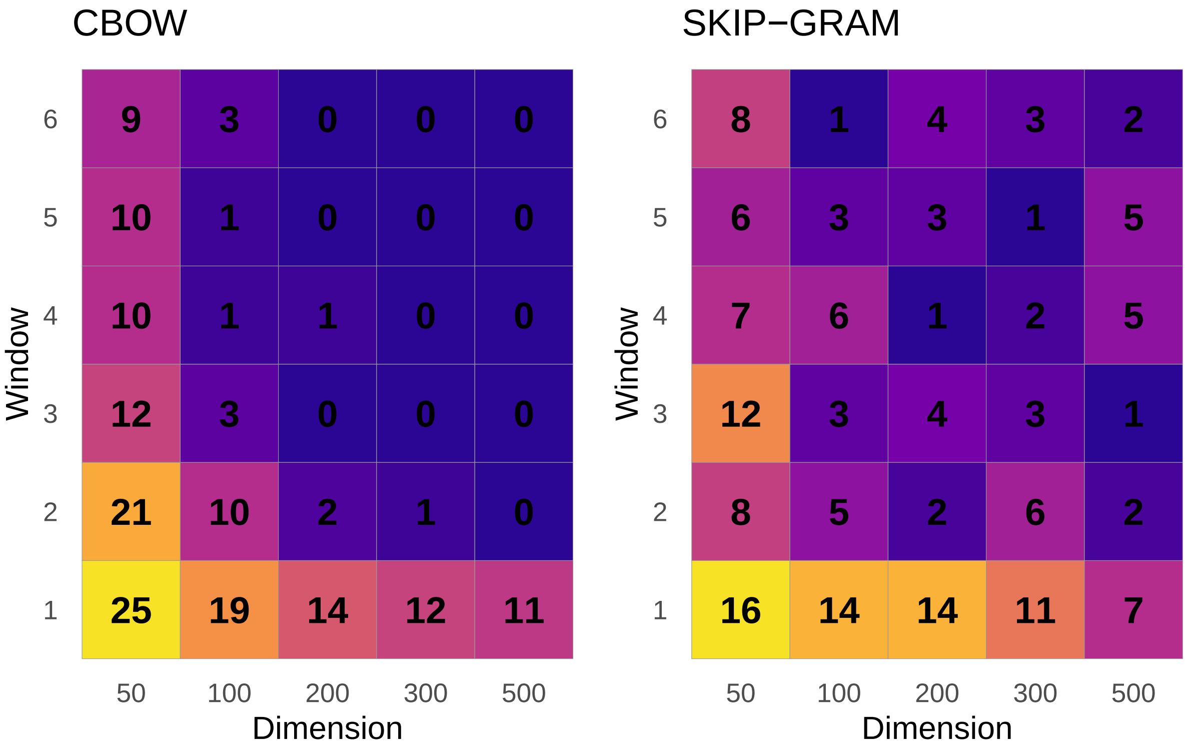

Figure 4. Heatmaps showing the number of languages for which each combination of embedding dimension (x-axis) and context window size (y-axis) produced the top three highest adjusted R^2 when predicting Wikipedia-based word frequency norms.

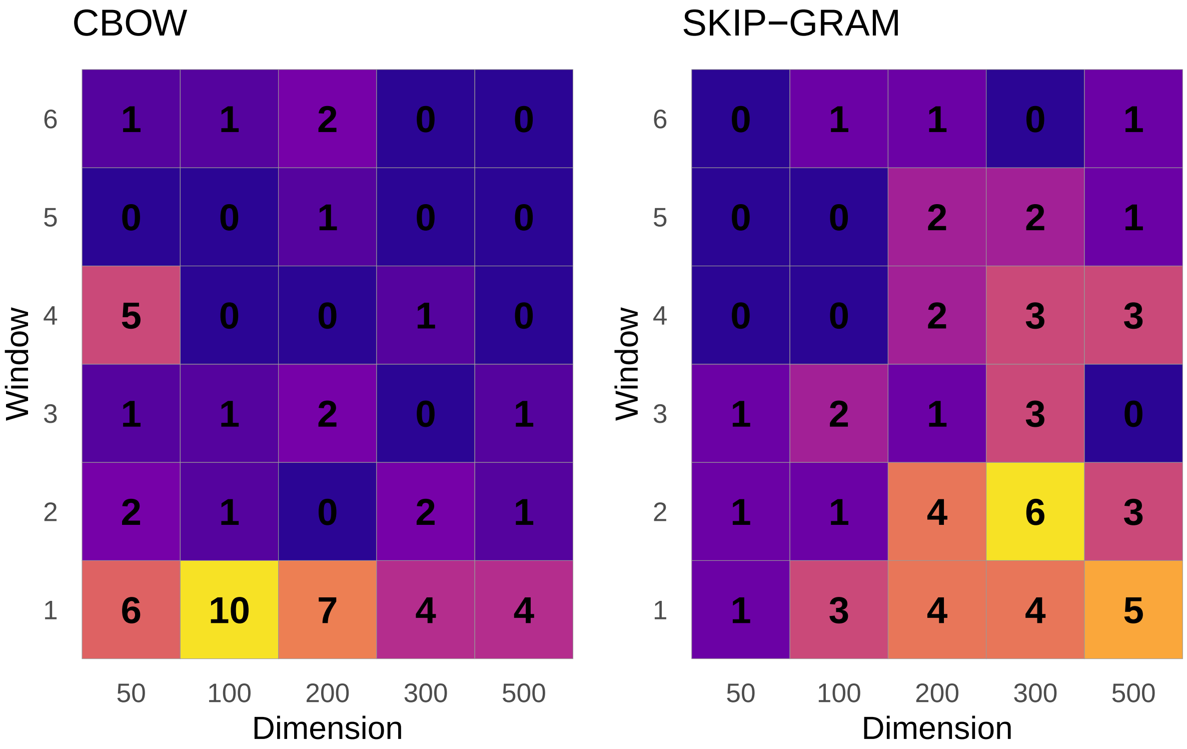

Figure 5. Heatmaps showing the number of languages for which each combination of embedding dimension (x-axis) and context window size (y-axis) produced the top three predictive values when predicting across an extended set of avaliable norms.

Appendix

Reproducibility

Manuscript

We used R version 4.4.2 (2024-10-31) to create this manuscript with the following packages:

[H]

Table A1

R packages and versions used in the analyses.

	package	loadedversion
dplyr	dplyr	1.1.4
ggplot2	ggplot2	3.5.2
ISOcodes	ISOcodes	2025.05.18
papaja	papaja	0.1.3
patchwork	patchwork	1.3.0
rio	rio	1.2.3
tidyr	tidyr	1.3.1
tinylabels	tinylabels	0.2.4
trackdown	trackdown	1.5.1

Note. Versions correspond to the computational environment at the time of knitting.