CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities

Viorica Marian, James Bartolotti, Sarah Chabal, and Anthony Shook*

Northwestern University

IN PRESS, PLoS ONE (2012)

Author Note

Viorica Marian, James Bartolotti, Sarah Chabal, and Anthony Shook, Department of Communication Sciences and Disorders, Northwestern University.

This work was funded by grant NICHD RO1 HD059858-01A to Viorica Marian. The authors thank members of the *Northwestern Bilingualism and Psycholinguistics Laboratory* for helpful comments on earlier drafts of this manuscript. We also thank the creators of the SUBTLEX databases, who provided the corpora used in the present study.

*All authors contributed equally to this work. Correspondence should be addressed to Viorica Marian, 2240 Campus Drive, Evanston, IL 60208 USA. Email: v-marian@northwestern.edu.

Abstract

Past research has demonstrated cross-linguistic, cross-modal, and task-dependent differences in neighborhood density effects, indicating a need to control for neighborhood variables when developing and interpreting research on language processing. The absence of a centralized database of neighborhood information has led to the inconsistent identification of neighbors, particularly across languages. The goals of the present paper are two-fold: (1) to introduce CLEARPOND (Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities), a centralized database of phonological and orthographic neighborhood information, both within and between languages, for five commonly-studied languages – Dutch, English, French, German, and Spanish; and (2) to show how CLEARPOND can be used to compare general properties of phonological and orthographic neighborhoods across languages in order to determine where and how languages differ in respect to their neighborhoods.

CLEARPOND allows researchers to input a word or list of words and obtain their phonological and orthographic neighbors, neighborhood densities, mean neighborhood frequencies, word lengths by number of phonemes and graphemes, and spoken-word frequencies. Neighbors can be defined by substitution, deletion, and/or addition, and the database can be queried separately along each metric or summed across all three. Neighborhood values can be obtained both within and across languages, and outputs can optionally be restricted to neighbors of higher frequency. To enable researchers to more quickly and easily develop stimuli, CLEARPOND can also search by features, generating lists of words that meet precise criteria, such as a specific range of neighborhood sizes, lexical frequencies, and/or word lengths.

CLEARPOND is freely-available to researchers and the public as a searchable, online database and for download at http://clearpond.northwestern.edu.

CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities

In research on language, neighborhoods are a conglomeration of words that are highly similar to one another along a critical characteristic. Most commonly, neighbors are defined on the basis of shared linguistic features such as orthography, phonology, or semantics. Because a word's neighborhood size (i.e., the number of neighbors it has; also called neighborhood density) can have an impact on a variety of linguistic tasks and processes, it has become an important psycholinguistic metric. However, in spite of the focus on neighbors in psycholinguistic research, neighbors are inconsistently identified, particularly across languages. These inconsistencies, which often arise as a result of researchers employing different databases, make it difficult to compare the effects of neighborhood density across studies. The current paper has two goals: (1) to introduce a centralized database of neighborhood information for five commonly-studied languages – Dutch, English, French, German, and Spanish – and provide a single corpus through which neighborhoods can be indexed cross-linguistically; and (2) to compare general properties of neighborhoods across these five languages using this database in order to determine where and how languages differ in respect to their neighborhoods.

In the current paper, we examined two types of linguistic neighborhoods – orthographic and phonological. Orthographic neighborhoods are often defined according to Coltheart, Davelaar, Jonasson, and Besner's [1] *N* metric, which refers to the number of words that can be constructed by substituting one letter of the target word. For example, the word *log* has *hog*, *lug*, and *lot* as orthographic neighbors. Phonological neighborhoods are calculated similarly, but instead of depending on grapheme substitution, phonological neighbors are constructed by substituting one phoneme of the target word [2]. *Fish* (/fɪʃ), for example, has *dish* (/dɪʃ/) and *fig*

(/fɪg/) as phonological neighbors. These "substitution neighbors" have historically been the focus of the literature and have dominated investigations of neighborhood size. However, research has also investigated the effects of addition (formed by the addition of a grapheme or phoneme, for example *and* has *hand* as an orthographic addition neighbor) and deletion (formed by the deletion of a grapheme or phoneme, for example *bend* has *end* as an orthographic deletion neighbor) neighbors [3].

The effects of phonological and orthographic neighborhood density on language processing have been well documented across a variety of tasks [4–11] and across multiple languages [12–15]. However, in spite of the prevalence of neighborhood effects, the nature of these effects is subject to debate. For example, neighborhood density may affect recognition and production processes differently [16,17], and effects may vary depending on the language of presentation [13,18,19] (but see [14]). The ongoing debate surrounding neighborhood density effects, particularly across languages, underscores the need for resources that allow researchers to consistently identify orthographic and phonological neighbors across studies. For some languages, even the most basic descriptive data are not available, forcing researchers to continually recreate basic neighborhood and frequency statistics. Furthermore, even when descriptive statistics are available [13,15,20,21], direct cross-linguistic comparisons are often not reported or possible.

While there have been some attempts to create consistent corpora from which neighborhood information can be derived, these corpora vary across languages. For example, N-Watch, a database of English neighborhood information [22], defines phonological neighbors according to the substitution of a single phoneme in any word position. BuscaPalabras, a database of Spanish neighborhood information [8], and E-Hitz, a database of Basque

neighborhood information [23], define phonological neighbors according to those same rules, but also include words that differ by the addition or deletion of a phoneme from any word position.

The goal of this paper is therefore to introduce CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities, a catalog of neighborhood density across languages. Perhaps the most comprehensive psycholinguistic database to date is WordGen [21], which queries the CELEX and Lexique databases to provide searchable datasets for Dutch, English, German, and French. While WordGen controls for factors such as written word frequency, orthographic neighborhood size, bigram frequency, and word length, it is missing a number of relevant features including information on phonological neighbors, neighborhood frequency, and the ability to index neighbors across languages. The database that we present here has been controlled for word frequency to ensure that consistent and comparable tokens are sampled from each language, and provides data regarding word length, neighborhood density, and neighborhood frequency. We also provide measures of foreign neighborhoods (i.e., the number of Spanish neighbors of an English word, or English neighbors of a Spanish word, etc.) for use in bilingual comparisons. Neighborhoods are defined both orthographically and phonologically, with stimuli derived from film and television subtitle corpora that capture spoken word frequencies. Finally, we have defined neighborhoods by substitution, addition, and deletion. It is our intent that CLEARPOND will provide a standard from which neighborhood data can be easily extracted and that it will provide a comprehensive tool for psycholinguistic researchers.

Methods

Selection of Corpora

To examine phonological and orthographic neighborhood densities across languages, we selected corpora for the following languages: Dutch (SUBTLEX-NL) [24], English (SUBTLEX-US) [25], French (Lexique) [26], German (SUBTLEX-DE) [27], and Spanish (SUBTLEX-ESP) [28]. Misspellings, including culturally-defined spellings (e.g., British "colour"), and foreign language intrusions (e.g., the English word "mind" in the Spanish corpus) were removed by cross-referencing each subtitle corpus with a dictionary in that language. Because all five corpora use the same source-material (i.e., film and television subtitles) to derive frequency data, they are highly comparable and well suited for cross-language comparisons. To increase similarity among the corpora, homographs were removed from the French corpus to match the parameters of the Dutch, English, German, and Spanish corpora (none of which distinguish between the different meanings of homographs). French homographs were reduced to a single entry, and the frequency per million of the collapsed entry was created by adding the frequency per million of each of the homographs. For example, the French word *est* is the third person singular form of the verb meaning "to be," and has a frequency of 19,417 per million; *est* is also the French word for the cardinal direction East, which has a frequency of 81 per million. We collapsed these two entries into a single entry, *est*, that had a frequency of 19,498 per million.

Using large corpora (the subtitle lexicons range from 74,286 to 441,132 tokens) can lead to overestimations of neighborhood size compared to people's actual working vocabularies. By only including words above a certain frequency threshold, the effect of very low frequency words (which are unlikely to be in people's everyday, working vocabularies) on neighborhood calculations is reduced. In the present study, a frequency threshold of 0.34 per million was used, based on the standard used by Davis[3] [22]. This frequency cutoff yielded a corpus size of 27,751 for English, which compares favorably to English vocabulary size estimates for educated adults

(20,000 word families) [29]. However, the frequency cutoff yielded different corpus sizes across languages (Dutch: $N = 31,691$; English: $N = 27,751$; French: $N = 34,113$; German: $N = 45,027$; Spanish: $N = 41,968$), which would limit our ability to make cross-linguistic comparisons. Larger corpora are likely to inflate neighborhood size estimates, as a larger overall sample pool results in a larger pool of potential neighbor-candidates. To alleviate this concern, corpus size was equated across languages by including the 27,751 most frequent words in each language (based on the smallest corpus, English) in all further comparisons. Figure 1a (left) shows that when corpus size was equated, the languages had comparable average frequencies (Dutch: 32.58, *SEM* = 3.10; English: 32.72, *SEM* = 3.18; French: 30.87, *SEM* = 2.64; German: 33.74, *SEM* = 2.74; Spanish: 33.87, *SEM* = 3.02), while Figure 1a (right) indicates that the languages differed in average frequency when corpus size was instead defined by a frequency threshold. In addition, frequency distributions (Figure 1b) were comparable across languages when corpus size was equated. Together, these results provide support for the ability to make direct comparisons between the size-equated corpora.

Insert Figure 1 Here

Calculating Neighborhoods

Orthographic neighborhoods. Orthographic neighbors consisted of words that differed only by the addition, deletion, or substitution of a single grapheme, as this method of calculating neighbors (including addition, deletion, and subtraction neighbors) provides a stronger metric of the lexical-level influence of neighborhood density than typical measures of substitution neighbors alone [3]. For example, the word *plant* has neighbors like *planet* (addition), *plan* (deletion), and *plank* (substitution). Likewise, the English word *chief* and the French word *chien* (meaning dog) are cross-linguistic orthographic neighbors because they differ only in the

substitution of a single grapheme, “n” for “f.” Accented vowels and the Spanish “ñ” were treated as separate graphemes; therefore, words such as the French *ou* (English: ‘or’) and *où* (English: ‘where’) were considered to be orthographic neighbors.

Phonological neighborhoods. Phonological transcriptions of each orthographic entry in the corpora were created using eSpeak (http://espeak.sourceforge.net/), an open-source text-to-speech software that provides IPA transcriptions for multiple languages[4]. With this method, the phonological transcriptions of the corpora used machine-readable phonetic symbols based on the International Phonetic Alphabet so that language-to-language neighborhood comparisons are viable[5]. For examples of words in each language that correspond to each phoneme, see Appendix A.

Phonological neighbors were composed of words that differed in the addition, deletion or substitution of a single phoneme [18,30]. For instance, the English word *dough* (/doʊ/) shares a neighborhood with words like *dome* (/doʊm/; addition), *owe* (/oʊ/; deletion), and *show* (/ʃoʊ/; substitution) in English. In addition, the English word *eel* (/il/) and the Spanish word *hilo* (/ilo/) are cross-linguistic neighbors by virtue of the deletion of the final phoneme /o/ in the Spanish word.

Because the same subtitle corpora were used to calculate both orthographic and phonological neighborhoods, qualitative comparisons can be made across neighborhood types.

Foreign neighborhoods. The same methods that were used to calculate orthographic and phonological neighborhoods within languages were used to calculate foreign neighbors. We calculated the Dutch, French, German, and Spanish neighbors of every English word, as well as the English neighbors of every Dutch, French, German, and Spanish word. For these analyses, the pool of candidate neighbors included all 27,751 words within the foreign language’s database.

Because these foreign neighborhoods were constructed using the same databases used to calculate within-language neighborhoods, foreign and within-language neighborhoods of each language can be easily compared.

Results

Orthographic Neighborhoods

Orthographic word length. Average word length (in graphemes) was calculated for all 27,751 words in each language and was 8.41 (*SD* = 2.79) for Dutch, 7.26 (*SD* = 2.28) for English, 7.85 (*SD* = 2.26) for French, 8.25 (*SD* = 2.86) for German, and 7.94 (*SD* = 2.24) for Spanish; $F(4,138750) = 879.66$, $p < 0.001$. Follow-up tests revealed that group differences were significant between all language pairs. The distribution of word lengths for each language is shown in Figure 2.

Insert Figure 2 Here

Orthographic neighborhood size. The number of within-language substitution, addition, and deletion neighbors was calculated for each word in each language. The mean neighborhood sizes are shown in Figure 3. An ANOVA with language and word length as factors revealed a significant effect of language on total orthographic neighborhood size, $F(4,138690) = 12.69$, $p < 0.0001$, a significant effect of word length $F(12,138690) = 9829.49$, $p < 0.0001$, and a significant language x word length interaction $F(48,138690) = 222.25$, $p < 0.0001$. Post-hoc comparisons on the estimated marginal means for language revealed that English words contained significantly more neighbors than words in Dutch, French, German, or Spanish (all p's < 0.05)[6].

While the effect of substitution neighbors on linguistic processing has long been studied, recent evidence suggests that addition and deletion neighbors affect word processing as well [3].

To best characterize the effect of orthographic neighbors on word processing, all further analyses will consider the sum total of substitution, deletion, and addition neighbors for each word.

Insert Figure 3 Here

Distribution of orthographic neighborhood densities. Figure 4 shows the distribution of neighborhood densities across languages. The percentage of words in each language with at least one orthographic neighbor was 55.3% for Dutch, 64.1% for English, 77.2% for French, 61.0% for German, and 74.7% for Spanish.

Insert Figure 4 Here

Orthographic neighborhood size by word length. Figure 5 shows the average neighborhood size of words in each language for each word length.

Insert Figure 5 Here

Orthographic neighborhood size by word frequency. In each language, all 27,751 words were divided into twenty equally spaced frequency bins, with each bin representing a 5% increment. For example, bin one represented the average orthographic neighborhood size of the top 5% most frequent words in the language while bin 20 represented the average orthographic neighborhood size of the least frequent 5% of words. The average orthographic neighborhood size for words in each of these frequency bins is provided in Figure 6.

Insert Figure 6 Here

Foreign orthographic neighbors. Foreign orthographic neighborhoods were calculated for each English word in Dutch, French, German, and Spanish, and for each Dutch, French, German, and Spanish word in English. Results revealed that 21.2% of English words had at least one Dutch neighbor, 31.7% had at least one French neighbor, 23.6% had at least one German neighbor, and 21.7% had at least one Spanish neighbor. 28.0% of Dutch words, 33.9% of French

words, 30.0% of German words, and 22.8% of Spanish words had at least one English neighbor. The effect of foreign neighbors on orthographic neighborhood size is provided in Table 1.

Insert Table 1 Here

For each word with at least one within-language or foreign neighbor, the relative proportion of neighbors to all of a word's neighbors was calculated. Mean proportions are provided in Figure 7.

Insert Figure 7 Here

Phonological Neighborhoods

Phonological word length. Average word length (in phonemes) was calculated for all 27,751 words in each language and was 7.48 (SD = 2.51) for Dutch, 6.09 (SD = 2.18) for English, 5.77 (SD = 1.93) for French, 7.14 (SD = 2.45) for German, and 7.84 (SD = 2.28) for Spanish; $F(4,138750) = 4284.86$, $p < 0.001$. Follow-up tests revealed that group differences were significant between all language pairs. The distribution of word lengths for each language is shown in Figure 8.

Insert Figure 8 Here

Phonological neighborhood size. The number of within-language substitution, addition, and deletion neighbors was calculated for each word in each language. The mean neighborhood sizes are shown in Figure 9. An ANOVA with language and word length as factors revealed a significant effect of language on total phonological neighborhood size, $F(4,138695) = 2730.64$, $p < 0.0001$, a significant effect of word length $F(11,138695) = 10204.84$, $p < 0.0001$, and a significant language x word length interaction $F(44,138695) = 913.84$, $p < 0.0001$. Post-hoc comparisons on the estimated marginal means for language revealed that all languages differed on phonological neighborhood size (all p's < 0.05). As in the orthographic neighborhood

analyses, all further phonological neighborhood analyses consider the total number of substitution, addition, and deletion neighbors for each word.

Insert Figure 9 Here

Distribution of phonological neighborhood densities. Figure 10 shows the distribution of phonological neighborhood densities across languages. The percentage of words in each language with at least one neighbor was 55.2% for Dutch, 69.1% for English, 75.5% for French, 61.9% for German, and 74.6% for Spanish.

Insert Figure 10 Here

Phonological neighborhood size by word length. Figure 11 shows the average neighborhood size in each language for each word length.

Insert Figure 11 Here

Phonological neighborhood size by word frequency. In each language, all 27,751 words were divided into twenty equally spaced frequency bins (as was done with orthographic neighborhoods). The average phonological neighborhood size for words in each frequency bin is provided in Figure 12.

Insert Figure 12 Here

Foreign phonological neighbors. Foreign phonological neighborhoods were calculated for each English word in Dutch, French, German, and Spanish, and for each Dutch, French, German, and Spanish word in English. Results revealed that 15.9% of English words had at least one Dutch neighbor, 10.6% had at least one French neighbor, 15.8% had at least one German neighbor, and 4.8% had at least one Spanish neighbor. 10.8% of Dutch words, 12.0% of French words, 12.4% of German words, and 1.6% of Spanish words had at least one English neighbor. The effect of foreign neighbors on phonological neighborhood size is provided in Table 2.

Insert Table 2 Here

For each word with at least one within-language or foreign neighbor, the relative proportion of within-language or foreign neighbors to all of a word's neighbors was calculated. Mean proportions are provided in Figure 13.

Insert Figure 13 Here

Discussion

The results of our analyses show consistent patterns across languages in the effects of word length and lexical frequency on neighborhood size. Differences across languages are also present – for example, while French has the most phonological neighbors, English contains more orthographic neighbors than the other four languages examined. The degree of similarity between phonological and orthographic neighbors also varies across languages (e.g., in Spanish, phonological and orthographic neighborhoods are more alike than in any other language). Within languages, differences emerge dependent on neighborhood origin; foreign neighbors are relatively infrequent compared to within-language neighbors.

Comparing Orthographic and Phonological Neighborhoods

Because the present analysis derived orthographic and phonological neighborhoods from the same subtitle corpora, we were able to make direct comparisons between the two neighborhood types. The differences that emerge in the relationships between these neighborhoods across languages can potentially be used to illuminate differences in language transparency. Transparency is a measure of how closely a language maintains a one-to-one grapheme-phoneme correspondence; the more transparent a language, the more the graphemes and phonemes are tightly matched. For example, in the most transparent of languages, each phoneme would map to only one grapheme and vice versa (e.g., the Spanish phoneme /i/ is

always represented by the grapheme *i*, and the *i* grapheme always corresponds to the phoneme /i/). Conversely, opaque languages are those in which grapheme-phoneme mappings are less consistent; multiple graphemes can represent the same phoneme (e.g., English *k* and *c* can both represent the phoneme /k/), and more than one phoneme may be represented by a single grapheme (e.g., English *g* can represent the phonemes /g/ and / dʒ/). Because the grapheme-phoneme mappings of transparent languages are consistent, it is intuitive that, in these languages, many orthographic neighbors are also phonological neighbors. When phonemes and graphemes are consistently matched, the phonetic transcriptions of words mirror the orthographic structure. Therefore, when a single *grapheme* substitution (or addition or deletion) results in the creation of a new word, it is likely that the new word similarly differs from the original in only one *phoneme*. The more consistent the grapheme-phoneme mapping of a particular language, the more transparent the writing system.

Our analyses suggest that, in addition to indexing language transparency as a strict match between grapheme-phoneme correspondences, there may be a relationship between a language's transparency and the degree of similarity between the language's orthographic and phonological neighborhoods. For example, Spanish and German (both considered to be transparent languages [31]), demonstrate a high degree of similarity in the distributions of their orthographic and phonological neighborhoods. However, the similarity between orthographic and phonological neighborhoods is not quite as tightly coupled in German as it is in Spanish, likely because, German contains specific consonant clusters (e.g., *sch*) that correspond to single phonemes (e.g., /ʃ/). Accordingly, there is higher similarity between graphemic and phonemic word lengths in Spanish than in German, Dutch, or English (Figure 14). French, a language with a high number

of silent letters and digraphs, has the largest difference between graphemic and phonemic word length.

Insert Figure 14 Here

Comparing Types of Neighbors

In addition to revealing differences between phonological and orthographic neighborhoods, our data illustrate differences in how substitution, addition, and deletion neighbors are used across languages.

Orthographic neighborhoods. Relative to the other four languages, English contains a large number of orthographic substitution neighbors. This suggests that English makes use of more available letter sequences at every word length, and efficiently uses its graphemic space. In contrast, French derives a greater percentage of its neighbors from addition and deletion relative to the other languages. Although French has relatively few substitution neighbors, it nevertheless has the second largest total number of neighbors; this is driven by French's increased use of addition and deletion neighbors.

Phonological neighborhoods. A notable trend that emerged in the comparison of phonological neighborhood sizes across languages is the much higher occurrence of phonological neighbors of all types (substitution, deletion, and addition) in French when compared to all other languages. One potential explanation for the observed trend is the large number of homophones in the French language.

Homophones increase the phonological neighborhood density of a language because there are multiple lexical entries with the same phonological make-up. Therefore, if a word has a phonological neighbor that is one meaning of a homophonic word set, it also automatically has a phonological neighbor comprised of all other homophones. In languages such as French, where

homophonic word sets are numerous, the phonetic diversity of all tokens is decreased, and the pool of potential phonological neighbors is increased. For example, the French word *mer* (sea) is a substitution neighbor of *ver* (earthworm), *vers* (towards), *vert* (green), and *verre* (drinking glass), which are all pronounced /vɛʁ/; only *ver* would be an orthographic neighbor. The homophone account of French's increased phonological neighborhood density is consistent with an analysis of phonetic diversity across languages: French only contained 17,303 unique phonetic words (out of 27,751; 62.4%), compared to 27,258 in Dutch (98.0%), 27,007 in English (97.3%), 27,284 in German (98.3%), and 27,101 in Spanish (97.7%).

Foreign Neighborhoods

In our analysis of foreign neighbors, we restricted comparisons to English and each other language (Dutch, French, German, and Spanish) to facilitate ease of comparisons, and because English is one of the most commonly learned second languages [32]. Foreign orthographic neighbors were found to make relatively substantial contributions to overall neighborhood size, constituting between 13-20% of a word's total neighbors on average. Within-language neighbors still dominated overall neighborhood size, likely because languages have different orthotactic rules and requirements for the formulation of valid words. The result is that words in each of the languages we examined were more similar in orthographic form to other words within the same language than they were to foreign words.

Compared to foreign orthographic neighbors, foreign phonological neighbors were very rare. The effect of foreign phonological neighbors on overall neighborhood size was quite low, and the percentage of a word's neighbors that derived from a foreign language was even lower, between 1-8%. These results are consistent with those of Vitevitch [30], who conducted an

analysis of foreign phonological neighbors across Spanish and English and found that the two languages share relatively few neighbors.

One potential reason for the small number of foreign neighbors is that though the five languages we investigated share an alphabetic system (aside from accented letters), they contain phonological systems that are much more distinct. Because the orthographic structure of a language is anchored by that language's writing system, orthography does not vary much over time. Conversely, a language's phonetic structure has much more freedom to vary over time and across geographical space; the accumulation of these phonological changes likely contributes to the languages' phonological distinctiveness, thereby reducing the number of foreign phonological neighbors.

While comparisons of foreign neighbors can be used for purposes of stimuli construction and to validate cross-linguistic comparisons, it is important to note that our data should not be interpreted as a measure of the bilingual mental lexicon. In order to make true claims about the nature of bilingual lexical representations based on corpus analyses, it would first be necessary to procure a bilingual corpus in which frequency values are representative of usage when a single individual speaks two languages. To our knowledge, such a corpus does not exist[7]. If bilingual corpora can be obtained, it would be worthwhile to conduct neighborhood analyses using those lexical entries.

Conclusions and Future Directions

The corpus analysis presented in the current study provides a novel tool for researchers who study language processing. It enables comparisons between orthographic and phonological neighbors and within and across five languages.

While neighborhood information for some languages has been made available in the past [13,15,20,21], the database that we present here provides comparable corpora and analyses across languages. We also expand upon the past examinations of foreign neighbors in Spanish and English [30] by supplying foreign neighborhood data for four language pairs – English-Dutch, English-French, English-German, English-Spanish – and by including both orthographic and phonological neighbors. Our future efforts will focus on developing a comparable corpus derived from written word data using written-word databases, such as Google Ngram (http://books.google.com/ngram) to complement our present work on spoken language.

In sum, the current paper presents a unified database for indexing neighborhood information derived from spoken corpora. These data provide cross-linguistic metrics that are crucial for designing experiments of spoken and written language processing. We have made our database available in searchable form (see Appendix B for a description of the web interface) at http://clearpond.northwestern.edu; it is also freely available for download.

References

1. Coltheart M, Davelaar E, Jonasson JT, Besner D (1977) Access to the internal lexicon. Attention and Performance VI. pp. 535–555.
2. Luce PA, Pisoni D, Goldinger S (1990) Similarity neighborhoods of spoken words. Cognitve Models of Speech Processing: Psycholinguistic and Computational Perspectives. pp. 122–147.
3. Davis CJ, Perea M, Acha J (2009) Re(de)fining the orthographic neighborhood: The role of addition and deletion neighbors in lexical decision and reading. Journal of Experimental Psychology: Human Perception and Performance 35: 1550–1570. doi:10.1037/a0014253.
4. Siakaluk PD, Sears CR, Lupker SJ (2002) Orthographic neighborhood effects in lexical decision: The effects of nonword orthographic neighborhood size. Journal of Experimental Psychology: Human Perception and Performance 28: 661–681. doi:10.1037//0096-1523.28.3.661.
5. Andrews S (1989) Frequency and neighborhood effects on lexical access: Activation or search? Journal of Experimental Psychology: Learning, Memory, and Cognition 15: 802–814. doi:10.1037//0278-7393.15.5.802.
6. Tsai J, Lee C, Lin Y, Tzeng O, Hung D (2006) Neighborhood size effects of Chinese words in lexical decision and reading. Word Journal of the International Linguistic Association: 659–675.
7. Andrews S (1997) The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. Psychonomic Bulletin & Review 4: 439–461.

8. Davis CJ, Perea M (2005) BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. Behavior Research Methods 37: 665–671. doi:10.3758/BF03192738.

9. Yates M (2005) Phonological neighbors speed visual word processing: Evidence from multiple tasks. Journal of Experimental Psychology: Learning, Memory, and Cognition 31: 1385–1397. doi:10.1037/0278-7393.31.6.1385.

10. Grainger J, Muneaux M, Farioli F, Ziegler JC (2005) Effects of phonological and orthographic neighbourhood density interact in visual word recognition. The Quarterly Journal of Experimental Psychology 58: 981–998. doi:10.1080/02724980443000386.

11. Baese-Berk M, Goldrick M (2009) Mechanisms of interaction in speech production. Language and Cognitive Processes 24: 527–554. doi:10.1080/01690960802299378.

12. Marian V, Blumenfeld HK, Boukrina OV (2008) Sensitivity to phonological similarity within and across languages. Journal of Psycholinguistic Research 37: 141–170. doi:10.1007/s10936-007-9064-9.

13. Vitevitch MS, Rodríguez E (2005) Neighborhood density effects in spoken word recognition in Spanish. Journal of Multilingual Communication Disorders 3: 64–73. doi:10.1080/14769670400027332.

14. Baus C, Costa A, Carreiras M (2008) Neighbourhood density and frequency effects in speech production: A case for interactivity. Language and Cognitive Processes 23: 866–888. doi:10.1080/01690960801962372.

15. Frauenfelder UH, Baayen RH, Hellwig FM, Schreuder R (1993) Neighborhood density and frequency across languages and modalities. Journal of Memory and Language 32: 781–804. doi:10.1006/jmla.1993.1039.

16. Dell G (2003) Neighbors in the lexicon: Friends or foes? In: Schiller N, Meyer AS, editors. Phonetics and Phonology in Language. New York: Mouton De Gruyter. pp. 9–47.

17. Gahl S, Yao Y, Johnson K (2012) Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. Journal of Memory and Language. doi:10.1016/j.jml.2011.11.006.

18. Luce PA, Pisoni DB (1998) Recognizing spoken words: The neighborhood activation model. Ear and Hearing 19: 1–36. doi:10.1097/00003446-199802000-00001.

19. Vitevitch MS, Stamer MK (2006) The curious case of competition in Spanish speech production. Language and Cognitive Processes 21: 760–770. doi:doi:10.1080/01690960500287196.

20. Balota DA, Yap MJ, Cortese MJ, Hutchison KA, Kessler B, et al. (2007) The English lexicon project. Behavior Research Methods 39: 445–459. doi:10.3758/BF03193014.

21. Duyck W, Desmet T, Verbeke LPC, Brysbaert M (2004) WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. Behavior Research Methods, Instruments, & Computers 36: 488–499. doi:10.3758/BF03195595.

22. Davis CJ (2005) N-watch: A program for deriving neighborhood size and other psycholinguistic statistics. Behavior Research Methods 37: 65–70. doi:10.3758/BF03206399.

23. Perea M, Urkia M, Davis CJ, Agirre A, Laseka E, et al. (2006) E-Hitz: W word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). Behavior Research Methods 38: 610–615.

24. Keuleers E, Brysbaert M, New B (2010) SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. Behavior Research Methods 42: 643–650. doi:10.3758/BRM.42.3.643.

25. Brysbaert M, New B (2009) Moving beyond Kucera 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: 977–990. doi:10.3758/BRM.41.4.977.

26. New B, Pallier C, Brysbaert M, Ferrand L (2004) Lexique 2: A new French lexical database. Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc 36: 516–524.

27. Brysbaert M, Buchmeier M, Conrad M, Jacobs AM, Bolte J, et al. (n.d.) The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. Experimental Psychology.

28. Cuetos F, Glez-Nosti M, Barbón A, Brysbaert M, Barbon A (2011) SUBTLEX-ESP: Spanish word frequencies based on film subtitles. Psicologica 32: 133–143.

29. Goulden R, Nation P, Read J (1990) How large can a receptive vocabulary be? Applied Linguistics 11: 341–363. doi:10.1093/applin/11.4.341.

30. Vitevitch MS (2012) What do foreign neighbors say about the mental lexicon? Bilingualism: Language and Cognition: 1–6. doi:10.1017/S1366728911000149.

31. Seymour PHK, Aro M, Erskine JM (2003) Foundation literacy acquisition in European orthographies. British Journal of Psychology 94: 143–174. doi:10.1348/000712603321661859.

32. Mejer L, Boateng S, Turchetti P (2010) More students study foreign languages in Europe but perceptions of skill levels differ significantly. Statistics in Focus 49.

33. Baayen RH, Piepenbrock R, Gulikers L (1995) The CELEX Lexical Database (Release 2) [CD-ROM].

34. Weide RL (1998) CMU pronunciation dictionary, release 0.6.

35. Rafalovitch A, Dale R (2009) United Nations general assembly resolutions: A six-language parallel corpus. Proceedings of the MT Summit.

Appendix A

All words were phonetically transcribed according to the IPA by the speech software eSpeak. The complete list of phonemes used is provided in Tables A1 (consonants) and A2 (vowels). Example words are provided for each language in which the phoneme appears, with the relevant letters highlighted in bold text.

Table A1 – Consonants and example words

IPA	Dutch	English	French	German	Spanish
Consonants					
b	he**bb**en	**b**ut	**b**ien	ha**b**en	**b**año
β	-	-	-	-	fa**v**or
ç	-	-	-	i**ch**	-
d	**d**at	o**dd**	**d**eux	**d**as	**d**el
dʒ	bu**dg**et	**j**ob	léthar**g**ie	mana**g**er	-
f	hee**f**t	**f**or	**f**aire	**v**on	**f**avor
g	zo**g**enaamde	**g**et	**g**rand	sa**g**en	**g**racias
ɣ	**g**een	-	-	-	se**g**uir
h	**h**ij	**h**at	-	**h**ast	-
j	**j**aar	**y**ou	fi**ll**e	**j**a	d**io**s
k	i**k**	**c**an	**qu**i	**k**ann	**qu**e
l	we**l**	**l**ike	**l**oup	**l**eute	fe**l**iz
ʎ	-	-	-	-	si**ll**a
m	**m**aar	**m**e	**m**ais	**m**it	**m**ucho
n	**n**ek	**n**eed	**n**ous	**n**icht	**n**ada
ŋ	jo**ng**	goi**ng**	parki**ng**	la**ng**	ci**n**co
ɲ	ora**ngj**e	-	si**gn**e	-	se**ñ**or
p	**p**raat	**p**ut	**p**our	**p**roblem	**p**ara
pf	-	-	-	**pf**erd	-
r	p**r**aat	-	-	-	gue**rr**a
ɾ	-	-	-	-	p**r**oblema
ʁ	-	-	t**r**ès	f**r**au	-
ɹ	-	**r**ight	-	-	-
s	men**s**en	**s**ome	**s**uis	i**s**t	lo**s**
ʃ	so**ci**aal	**sh**e	**ch**ez	**sch**ön	-
t	he**t**	**t**ime	**t**out	mi**t**	**t**odos
ts	-	-	-	**z**u	-
tʃ	**ch**ecken	whi**ch**	mat**ch**	deu**tsch**e	mu**ch**o
θ	-	**th**ink	-	-	ha**c**er

ð	-	**th**at	-	-	na**d**a
v	**v**oor	**v**ery	a**v**ec	**w**as	-
ʊ	**w**aarom	-	-	-	-
w	be**w**aren	**wh**at	**ou**i	-	b**u**eno
x	to**ch**	lo**ch**	x	au**ch**	di**j**o
z	**z**ijn	ha**s**	be**s**oin	**s**ehr	-
ʒ	vi**s**ioen	plea**s**ure	**j**e	passa**g**iere	-

Table A2 – Vowels and example words

IPA	Dutch	English	French	German	Spanish
Vowels					
æ	-	c**a**t	-	-	-
a	d**a**t	-	p**a**s	w**a**s	m**á**s
aː	n**aa**r	-	-	-	-
ɑ	-	f**a**ther	-	-	-
e	-	-	all**er**	-	hombr**e**
ɛ	g**e**ld	g**e**t	c**e**tte	d**e**nn	p**e**ro
ɛː	-	-	-	sp**ä**ter	-
ə	onz**e**	comm**a**	p**e**tit	bitt**e**	-
i	n**i**et	n**ee**d	qu**i**	l**ie**be	s**í**
ɪ	d**i**t	w**i**th	-	m**i**t	-
Ï	-	an**y**thing	-	-	-
œ	l**eu**k	-	p**eu**r	k**ö**nnen	-
ø	-	-	v**eu**x	sch**ö**n	-
o	-	-	v**o**tre	**o**der	h**o**la
ɔ	t**o**ch	**o**ff	-	n**o**ch	-
ɔː	s**oo**rt	-	-	-	-
u	g**oe**d	y**ou**	v**ou**s	g**u**te	l**u**gar
ʊ	-	p**u**t	-	**u**nd	-
ʌ	-	b**u**t	-	-	-
y	b**uu**rt	-	t**u**	**ü**ber	-
yː	d**uw**	-	-	f**ü**r	-
ʏ	h**u**t	-	-	-	-
Dipthongs					
aɪ	-	m**y**	-	**ei**n	b**ai**le
aʊ	-	n**ow**	-	fr**au**	p**au**sa
eɪ	-	th**ey**	-	-	s**ei**s
ɛi	h**ij**	-	-	-	-
œy	t**ui**n	-	-	-	-
oʊ	b**oo**t	kn**ow**	-	-	-
ɔɪ	-	b**oy**	-	-	est**oy**
ɔʏ	-	-	-	l**eu**te	-
ʌu	j**ou**	-	-	-	-
Nasal Vowels					
ɑ̃	-	-	d**an**s	-	-
ɛ̃	-	-	bi**en**	-	-
œ̃	-	-	auc**un**	-	-
ɔ̃	-	-	n**on**	-	-
Rhotic Vowels					
ɝ	-	h**er**	-	-	-

ɚ	-	nev**er**	-	ab**er**	-
ɑɹ	-	c**ar**	-	-	-
ɛɹ	-	th**er**e	-	-	-
iəɹ	-	n**ear**	-	-	-
oɹ	-	f**or**	-	-	-
ʊɹ	-	t**our**	-	-	-

Appendix B

Insert Figure 15 Here

CLEARPOND provides a user-friendly, web-based interface for obtaining Dutch, English, French, German, and Spanish phonological and orthographic neighborhood densities (or, PONDs). The search function allows users to search for POND information in any of the five languages using single word queries or by providing full lists of words. CLEARPOND provides a number of important psycholinguistic measures, such as neighborhood density and neighborhood frequency, both for within-language neighbors and foreign-language neighbors. With user-controlled output selection, researchers can choose the output parameters that are most relevant. In addition to allowing users to acquire data for specific words, CLEARPOND can also search by features so that researchers can generate new lists of words that meet precise criteria, such as a specific range of neighborhood sizes or lexical frequency (as provided by the Subtlex databases). Furthermore, multiple filters can be applied simultaneously, providing greater control for stimuli creation. Users also have the option of exporting their results directly to a text file, making it easy to create downloadable documents containing pertinent psycholinguistic measures for all of their stimuli. In addition to the web-based interface, more comprehensive lists containing of all of the information provided by the database are available for download, so that the entire CLEARPOND database can be accessed offline.

Footnotes

[1] One important caveat to claims about the effects of neighborhood density is that neighborhood literature has been dominated by experiments carried out in English. As we will discuss in following sections, there is evidence that neighborhoods operate differently across languages. The trends outlined in this introduction may not generalize to non-English languages.

[2] As indexed by the number of entries in linguistic corpora.

[3] The N-Watch, a popular Neighborhood Density Database for English [22], includes all of the words from the CELEX English Frequency database above 0.34 per million. This frequency threshold was derived by comparing a dictionary of 65,013 words to the full 17.9 million word CELEX database, and excising any word that occurred fewer than seven times in the corpus, resulting in a minimum frequency value of 0.34 per million.

[4] Although the French database Lexique [26] includes phonological transcriptions, we generated phonological transcription for French using eSpeak as well, in order to maintain consistency across corpora.

[5] eSpeak allows for consistent transcriptions to be made across languages, which facilitates cross-linguistic neighborhood comparisons. To ensure the validity of eSpeak transcriptions, we selected a subset of words from each language that existed in both CLEARPOND and in a phonetic database for that language. We then calculated phonological neighborhoods (including substitution, addition, and deletion neighbors) for each word twice, once using the output provided by eSpeak and once using the output from the external database. The neighborhoods obtained by the two different metrics were very highly correlated:

Dutch eSpeak comparison with the CELEX database [33]: $N = 26{,}358$, $R = 0.94$, $p < 0.001$.

English eSpeak comparison with the CMU database [34]: $N = 26{,}474$, $R = 0.97$, $p < 0.001$.

French eSpeak comparison with the Lexique database [26]: $N = 27{,}751$, $R = 0.96$, $p < 0.001$. German eSpeak comparison with the CELEX database [33]: $N = 21{,}609$, $R = 0.93$, $p < 0.001$. Spanish eSpeak comparison with the Busca Palabras database [8]: $N = 10{,}978$, $R = 0.97$, $p < 0.001$.

[6] The longest 5% of all words were collapsed into a single category for analysis purposes. The magnitude of these results is likely driven, in part, by the large *N* values. Because every item from our database was analyzed, the values represent an entire population. Inferential statistics are more appropriate for drawing assumptions about a population from a relatively small sample.

[7] The United Nations produced a parallel corpus consisting of six languages' translations of 2100 United Nations General Assembly Resolution [35]. Parallel corpora, however, still fail to take into account the bilingual's potentially mixed lexicon.

Table 1

Mean orthographic within-language neighborhood size and foreign neighborhood size.

	Within-Language Neighborhood Size	Foreign Neighborhood Size				
		English	Dutch	French	German	Spanish
English	2.83		1.00	1.00	0.99	0.63
	(0.03)		(0.02)	(0.01)	(0.01)	(0.01)
Dutch	2.00	1.00				
	(0.02)	(0.02)				
French	2.35	1.00				
	(0.02)	(0.01)				
German	1.97	0.99				
	(0.02)	(0.01)				
Spanish	2.23	0.63				
	(0.02)	(0.01)				

Note. Values represent means, those in parentheses represent standard error of the mean.

Table 2

Mean phonological within-language and foreign neighborhood size

	Within-Language Neighborhood Size	Foreign Neighborhood Size				
		English	Dutch	French	German	Spanish
English	5.49		0.89	1.23	0.89	0.15
	(0.06)		(0.02)	(0.04)	(0.02)	(0.01)
Dutch	3.05	0.89				
	(0.04)	(0.02)				
French	10.32	1.23				
	(0.10)	(0.04)				
German	3.02	0.89				
	(0.03)	(0.02)				
Spanish	2.63	0.15				
	(0.02)	(0.01)				

Note. Values represent means, those in parentheses represent standard error of the mean.

Figure 1. (a) Word frequency (per million) across Dutch, English, French, German, and Spanish. Equating corpus sizes (left) resulted in average word frequencies that were comparable across languages; size-equated corpora were thus used in all further analyses. If, instead, corpus size was defined only by a frequency threshold (right), differences in average word frequency emerged. (b) Word frequency distributions for each language, using equivalent corpus sizes.

Figure 2. Distribution of orthographic word lengths for Dutch, English, French, German, and Spanish.

Figure 3. Mean orthographic neighborhood sizes for words in Dutch, English, French, German, and Spanish. Total mean neighborhood size (left group) includes single-letter substitutions (e.g., ‘log’ for ‘hog’), deletions (e.g., ‘end’ for ‘bend’) and additions (e.g., ‘hand’ for ‘and’).

Figure 4. Distribution of orthographic neighborhood densities across Dutch, English, French, German, and Spanish (log-log scale).

Figure 5. Average orthographic neighborhood size of words in Dutch, English, French, German, and Spanish at each word length.

Figure 6. Average orthographic neighborhood size as a function of word frequency. Frequency bins are evenly spaced divisions of words in 5% increments. Bin one represents the average orthographic neighborhood size of the top 5% most frequent words in the language, bin twenty represents the average orthographic neighborhood size of the 5% least frequent words.

Figure 7. Ratio of within-language and foreign orthographic neighbors as part of total neighborhood size for each word with at least one neighbor. The top row compares the proportion of English within-language neighbors (blue) to foreign neighbors in each other language. The bottom row compares the proportion of within-language neighbors in each language to foreign (i.e., English) neighbors (blue).

Figure 8. Distributions of phonological word lengths for Dutch, English, French, German, and Spanish.

Figure 9. Mean phonological neighborhood sizes for words in Dutch, English, French, German, and Spanish. Total mean neighborhood size (left group) includes single-phoneme substitutions (e.g., ‘show’ for ‘dough’), deletions (e.g., ‘owe’ for ‘dough) and additions (e.g., ‘dome’ for ‘dough).

Figure 10. Distribution of phonological neighborhood densities across Dutch, English, French, German, and Spanish (log-log scale).

Figure 11. Average phonological neighborhood size of words in Dutch, English, French, German, and Spanish at each word length.

Figure 12. Average phonological neighborhood size as a function of word frequency. Frequency bins are evenly spaced divisions of words in 5% increments. Bin one represents the average phonological neighborhood size of the top 5% most frequent words in the language, bin twenty represents the average phonological neighborhood size of the 5% least frequent words.

Figure 13. Ratio of within-language and foreign phonological neighbors as part of total neighborhood size for each word. The top row compares the proportion of English within-language neighbors (blue) to foreign neighbors in each other language. The bottom row compares the proportion of within-language neighbors in each language to foreign (i.e., English) neighbors (blue).

Figure 14. Comparisons of orthographic and phonological word lengths for Dutch, English, French, German, and Spanish.

Figure 15. Screen-shot of the EnglishPOND portion of the CLEARPOND website, accessible at http://clearpond.northwestern.edu.

Figure 1.

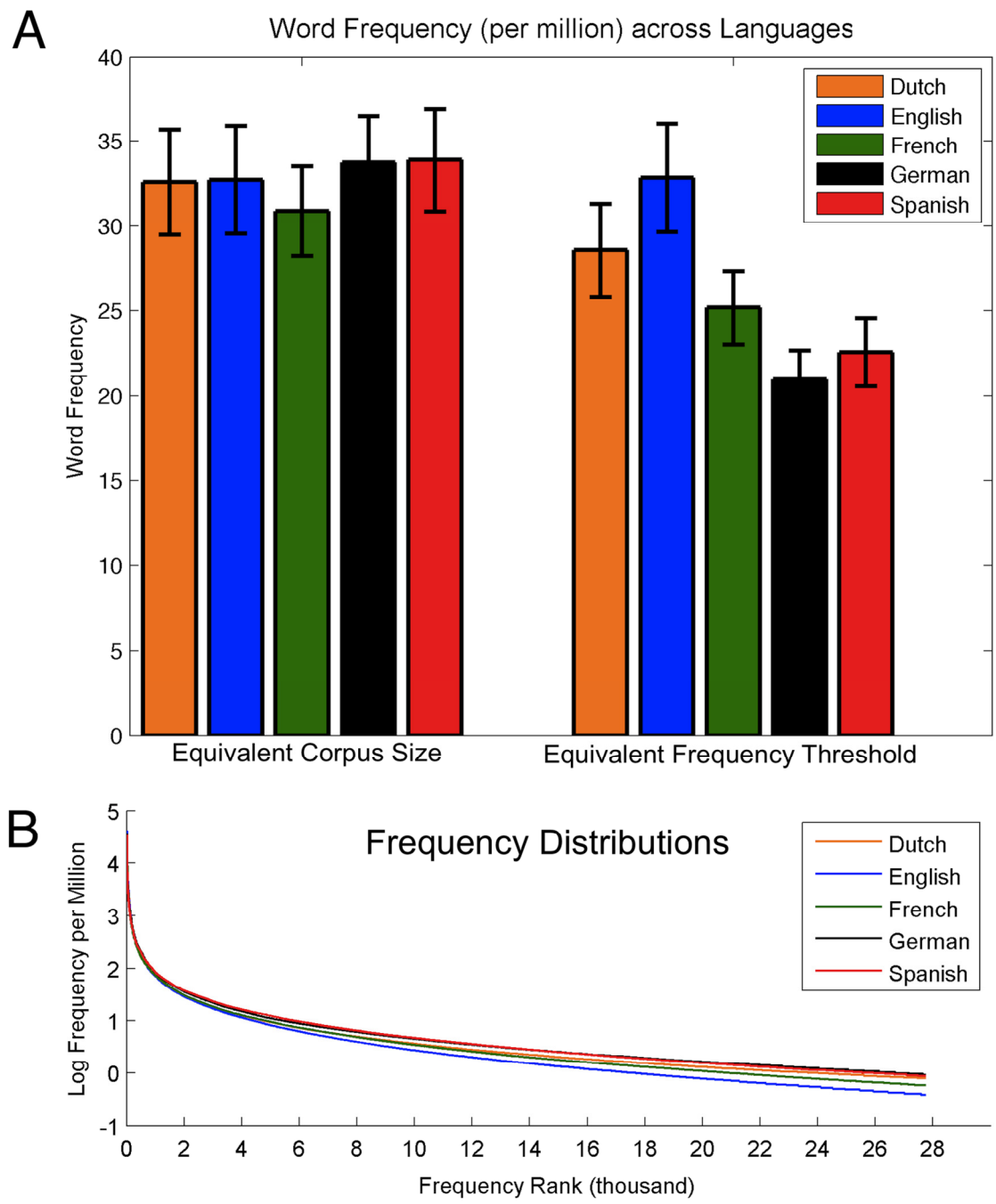

A
Word Frequency (per million) across Languages
Word Frequency
Dutch
English
French
German
Spanish
Equivalent Corpus Size
Equivalent Frequency Threshold
B
Frequency Distributions
Log Frequency per Million
Frequency Rank (thousand)

Figure 2.

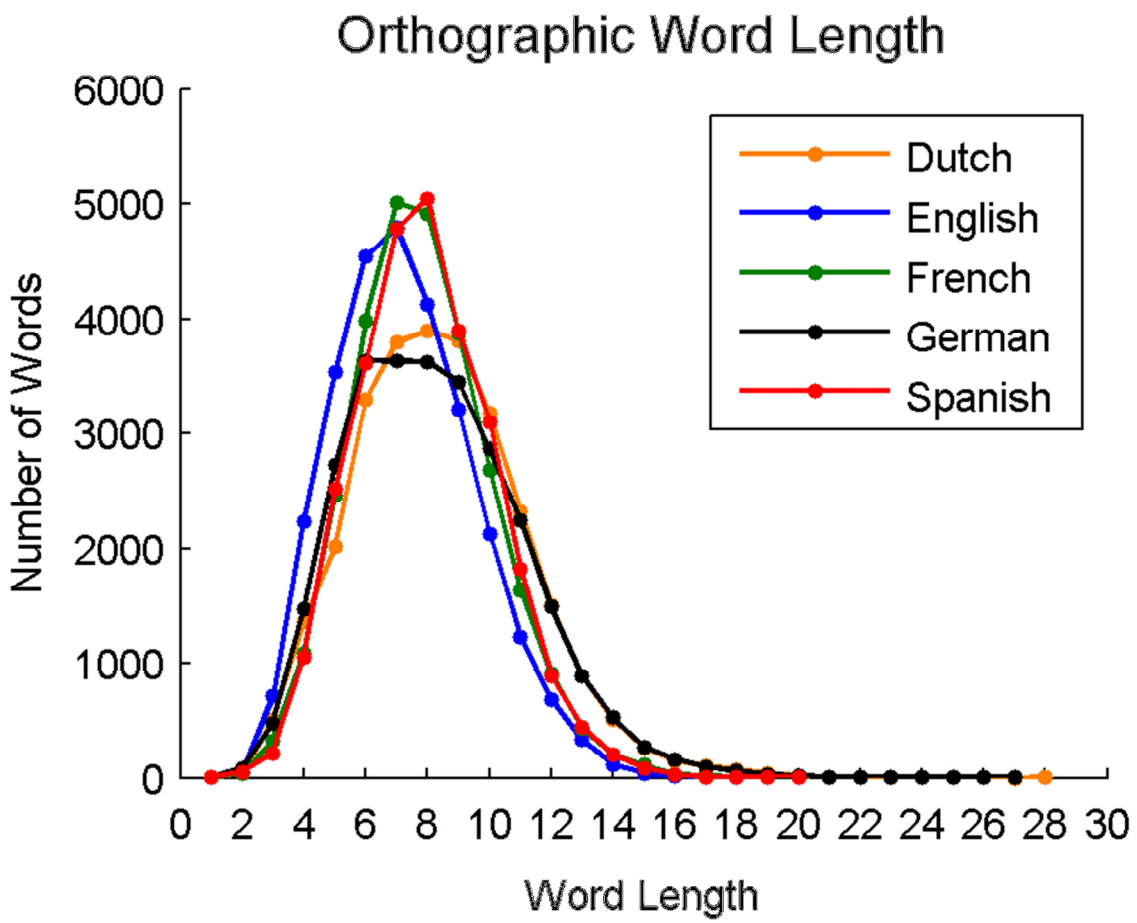

Orthographic Word Length
Number of Words
Word Length
0
1000
2000
3000
4000
5000
6000
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
Dutch
English
French
German
Spanish

Figure 3.

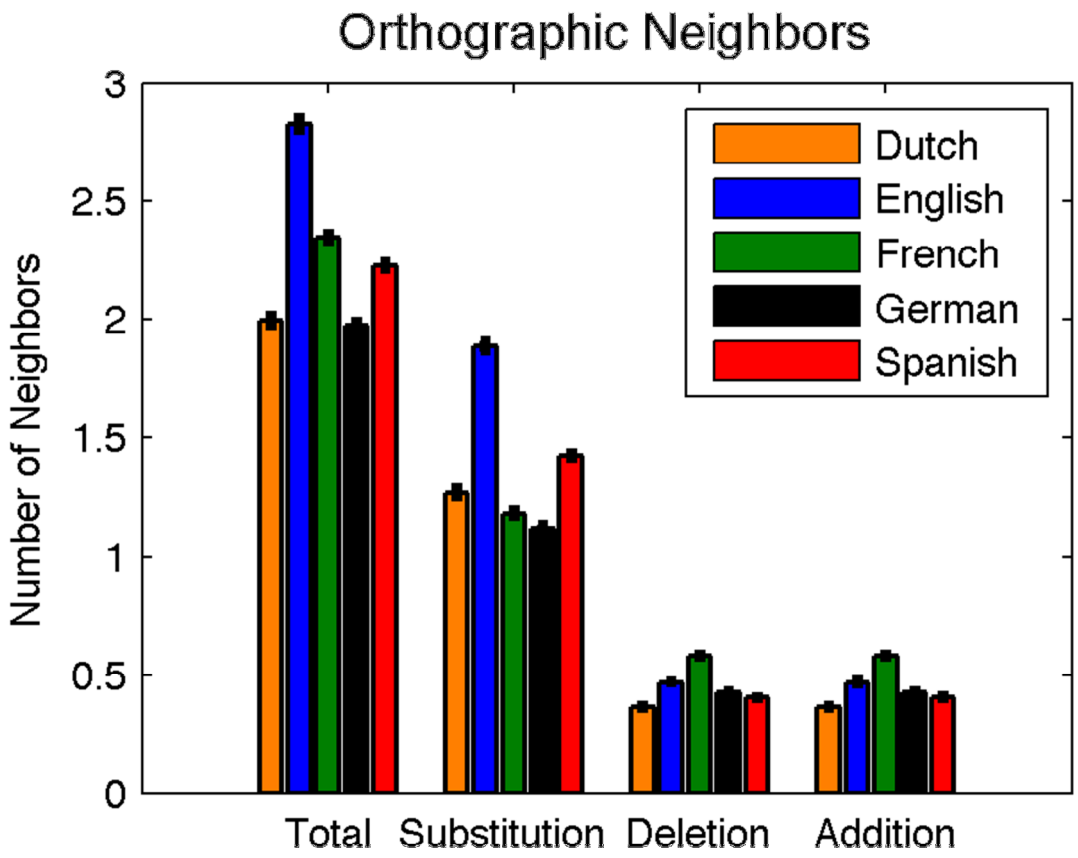

Orthographic Neighbors
Number of Neighbors
3
2.5
2
1.5
1
0.5
0
Dutch
English
French
German
Spanish
Total
Substitution
Deletion
Addition

Figure 4.

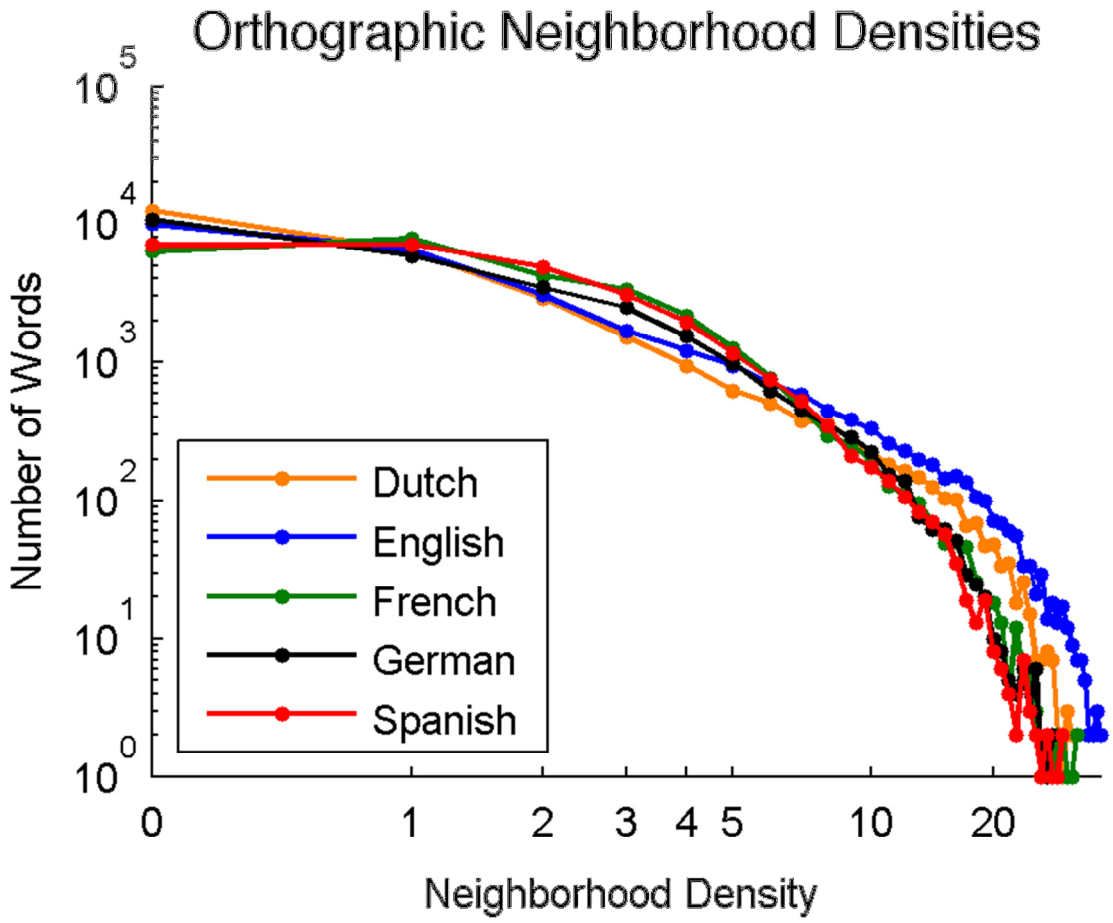

Orthographic Neighborhood Densities
Number of Words
Neighborhood Density
Dutch
English
French
German
Spanish
10^5
10^4
10^3
10^2
10^1
10^0
0
1
2
3
4
5
10
20

Figure 5.

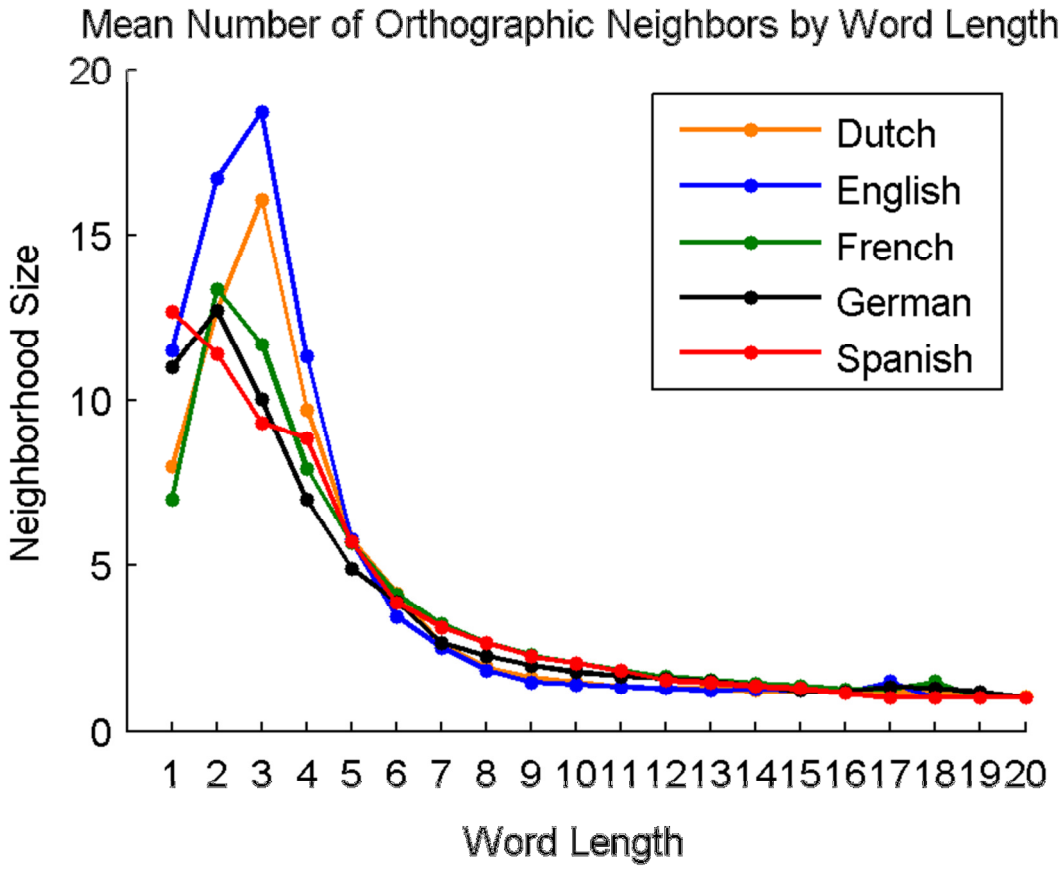

Mean Number of Orthographic Neighbors by Word Length
Neighborhood Size
Word Length
Dutch
English
French
German
Spanish
0
5
10
15
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Figure 6.

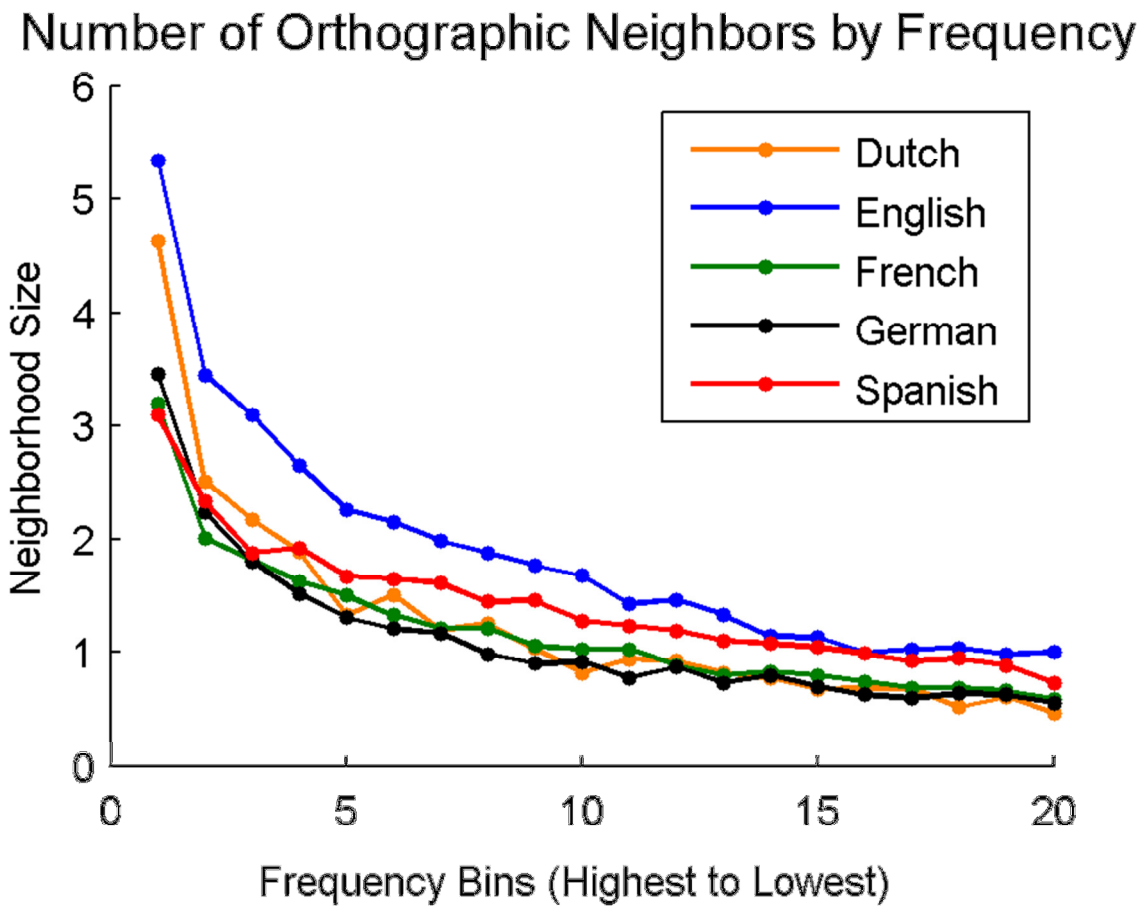

Number of Orthographic Neighbors by Frequency
Neighborhood Size
Frequency Bins (Highest to Lowest)
Dutch
English
French
German
Spanish
0
1
2
3
4
5
6
0
5
10
15
20

Figure 7.

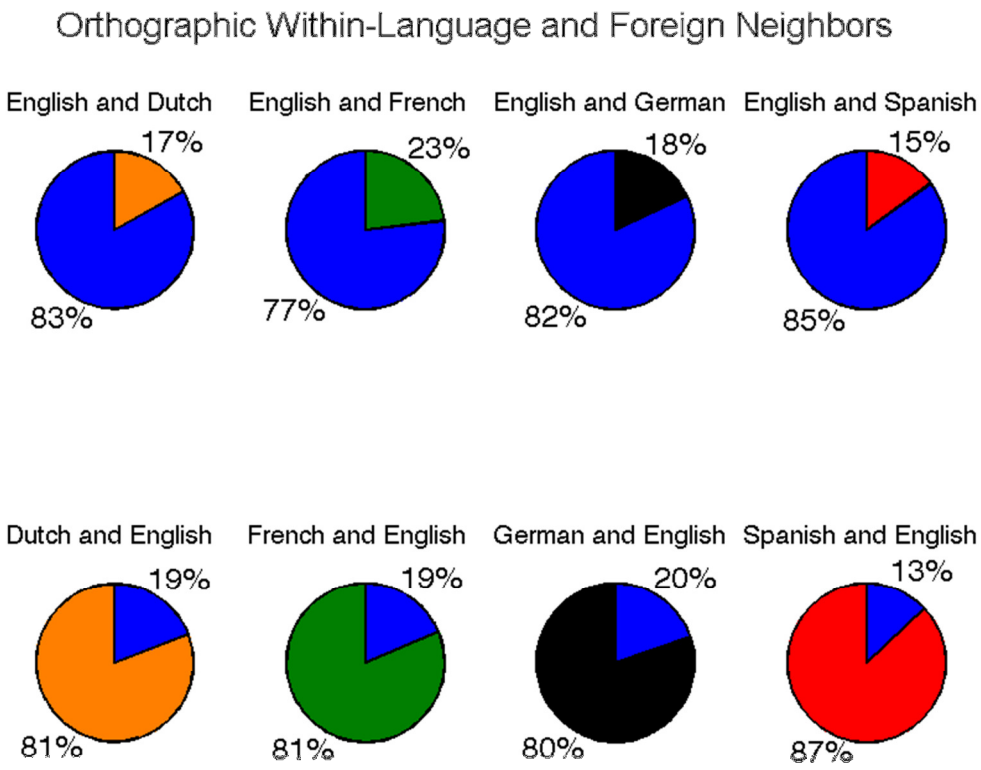

Orthographic Within-Language and Foreign Neighbors
English and Dutch
17%
83%
English and French
23%
77%
English and German
18%
82%
English and Spanish
15%
85%
Dutch and English
19%
81%
French and English
19%
81%
German and English
20%
80%
Spanish and English
13%
87%

Figure 8.

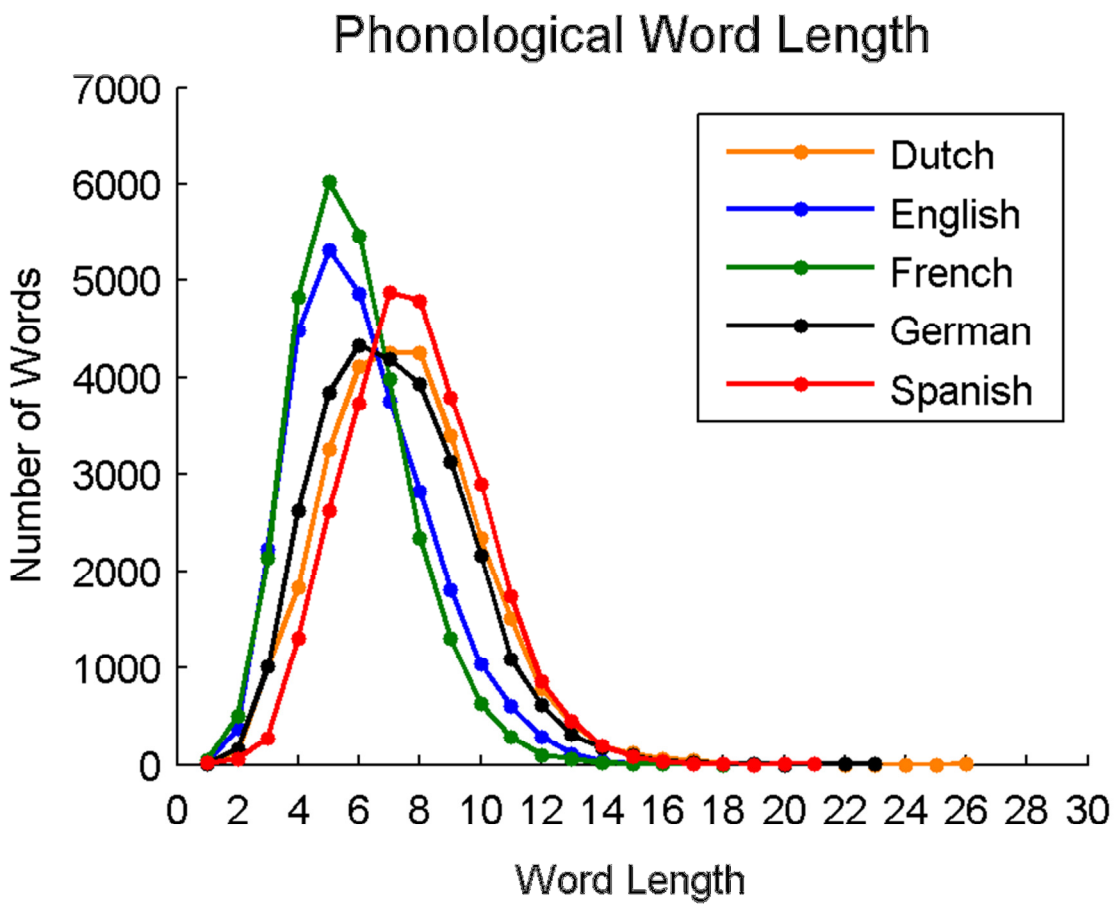

Phonological Word Length
Number of Words
Word Length
Dutch
English
French
German
Spanish
0
1000
2000
3000
4000
5000
6000
7000
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30

Figure 9.

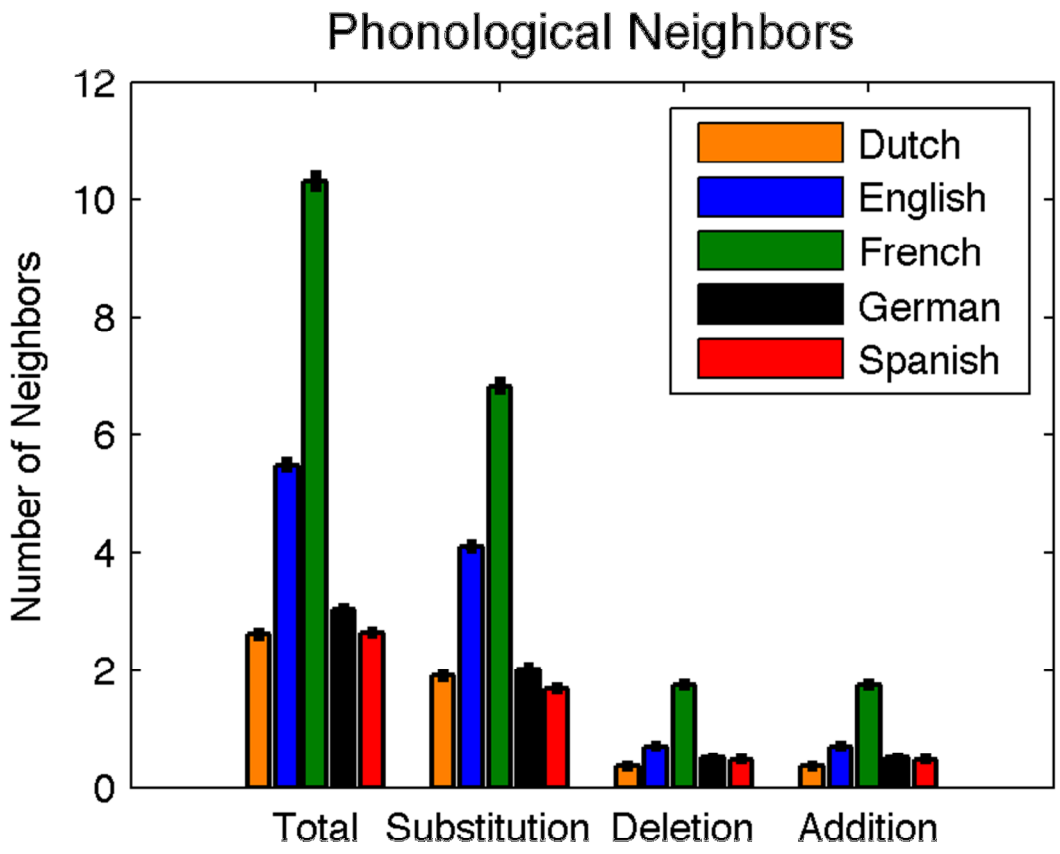

Phonological Neighbors
Number of Neighbors
0
2
4
6
8
10
12
Total
Substitution
Deletion
Addition
Dutch
English
French
German
Spanish

Figure 10.

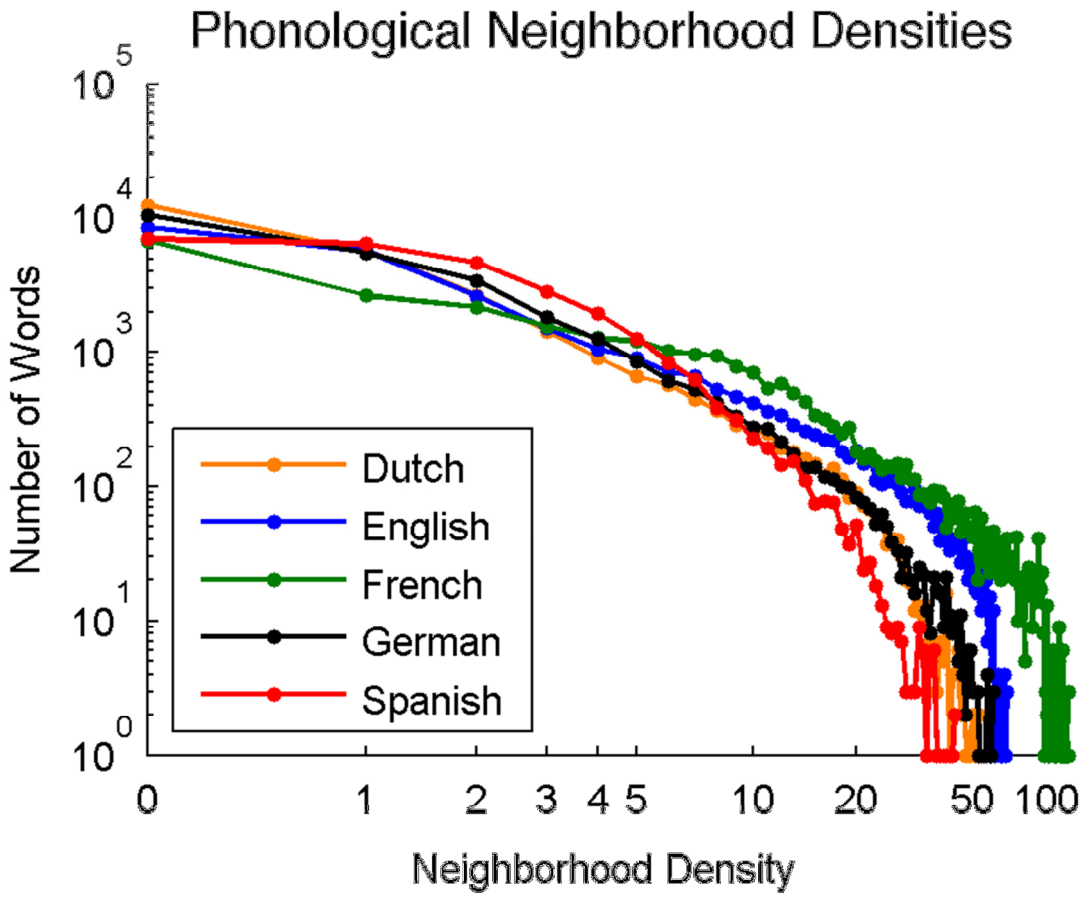

Phonological Neighborhood Densities
Number of Words
Neighborhood Density
Dutch
English
French
German
Spanish
10^5
10^4
10^3
10^2
10^1
10^0
0
1
2
3
4
5
10
20
50
100

Figure 11.

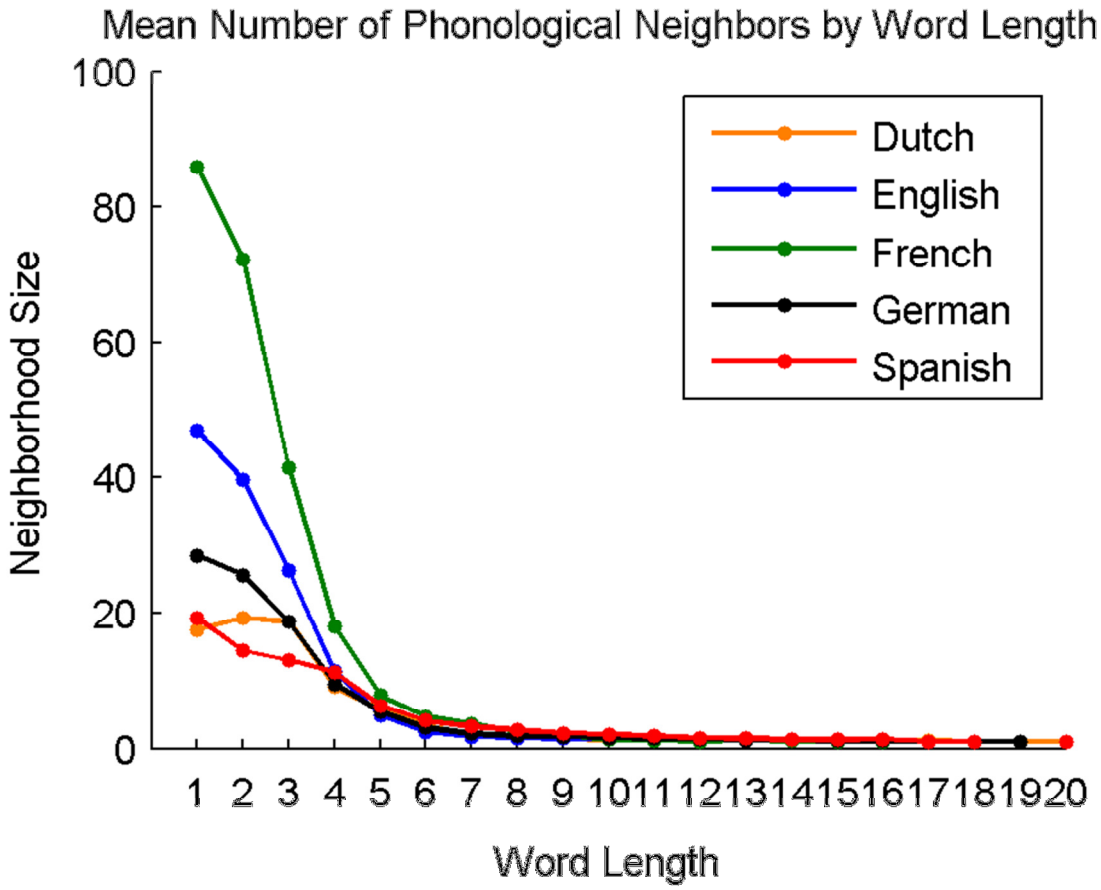

Mean Number of Phonological Neighbors by Word Length
Neighborhood Size
Word Length
Dutch
English
French
German
Spanish
0
20
40
60
80
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Figure 12.

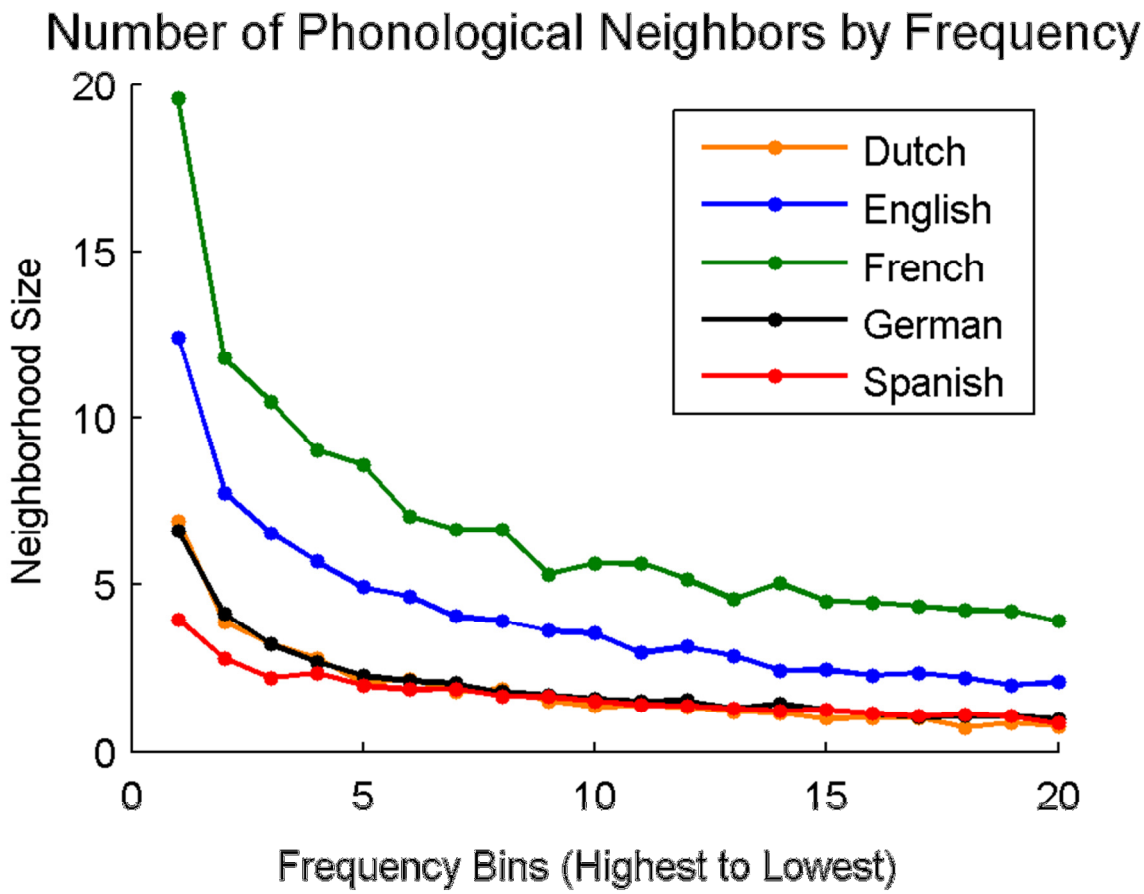

Number of Phonological Neighbors by Frequency
Neighborhood Size
Frequency Bins (Highest to Lowest)
0
5
10
15
20
Dutch
English
French
German
Spanish

Figure 13.

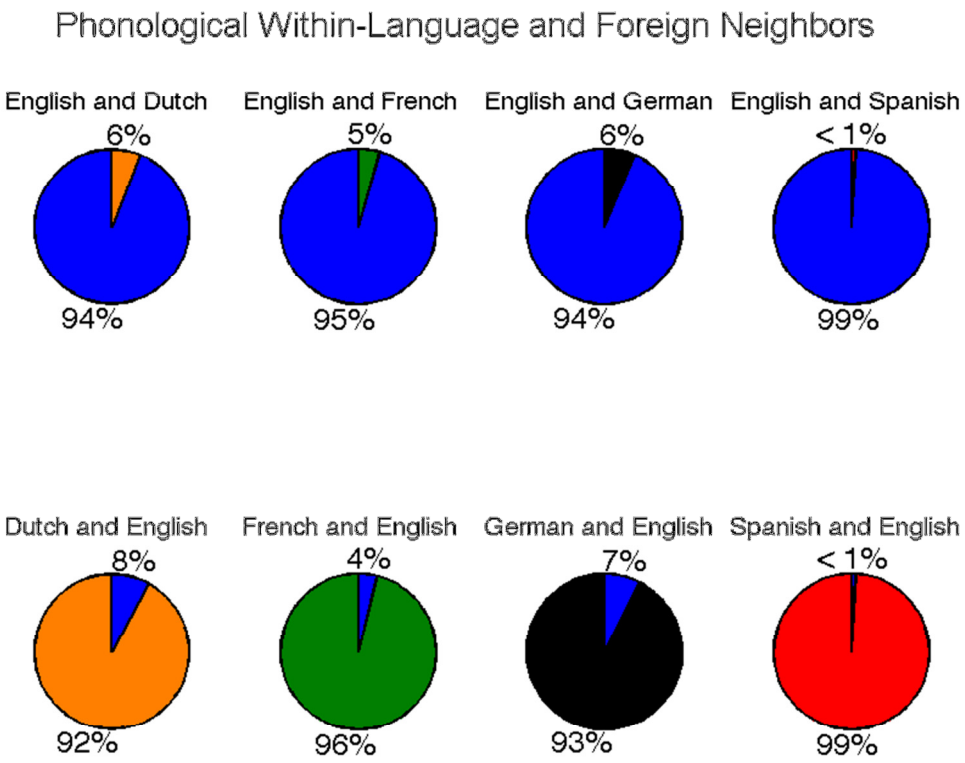

Phonological Within-Language and Foreign Neighbors
English and Dutch
6%
94%
English and French
5%
95%
English and German
6%
94%
English and Spanish
< 1%
99%
Dutch and English
8%
92%
French and English
4%
96%
German and English
7%
93%
Spanish and English
< 1%
99%

Figure 14.

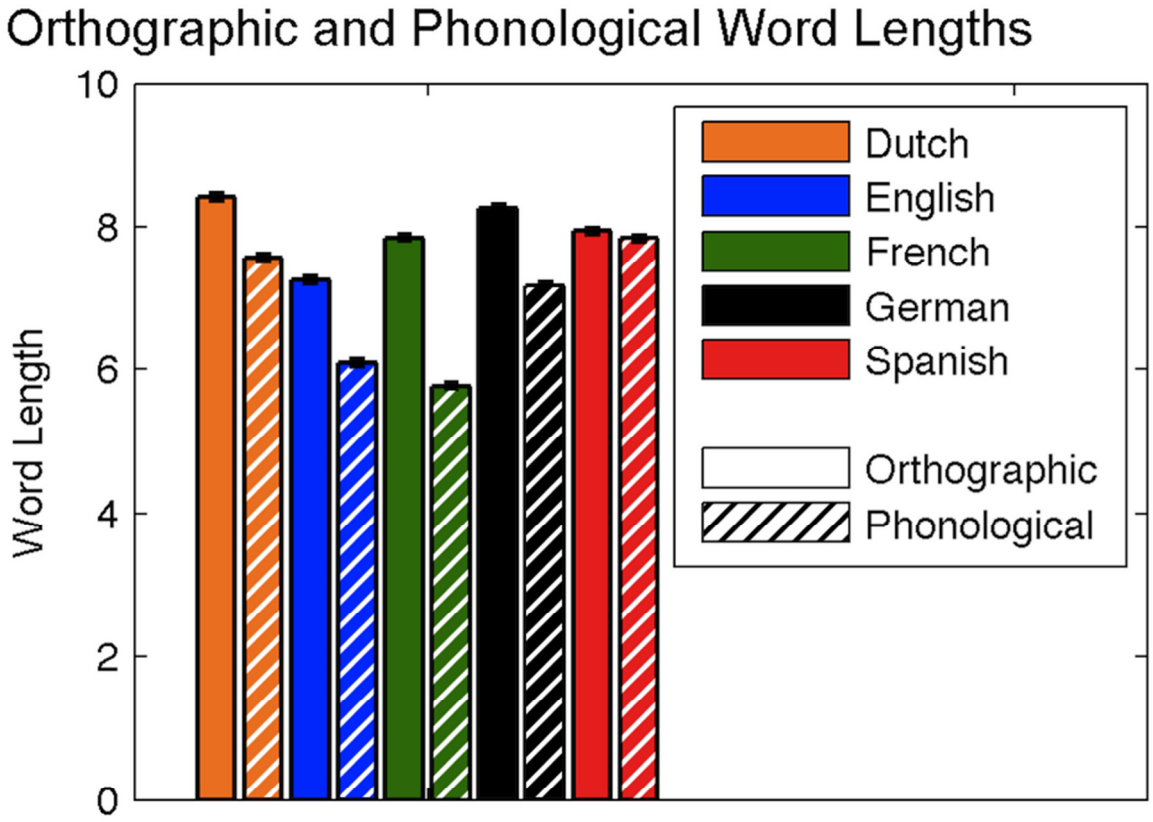
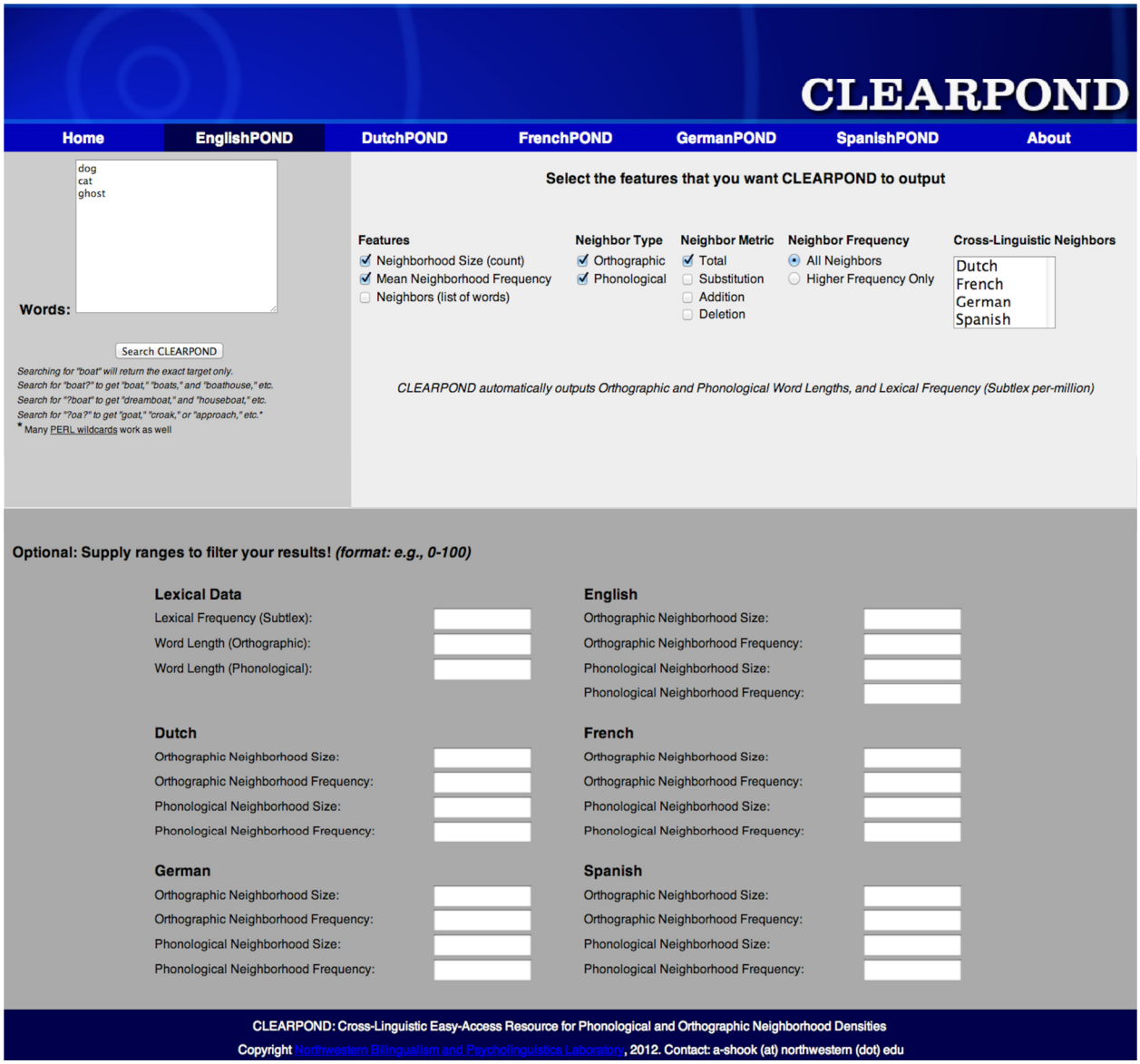

Orthographic and Phonological Word Lengths
Word Length
10
8
6
4
2
0
Dutch
English
French
German
Spanish
Orthographic
Phonological

Figure 15.


CLEARPOND
Home
EnglishPOND
DutchPOND
FrenchPOND
GermanPOND
SpanishPOND
About
Words:
dog
cat
ghost
Search CLEARPOND
Searching for "boat" will return the exact target only.
Search for "boat?" to get "boat," "boats," and "boathouse," etc.
Search for "?boat" to get "dreamboat," and "houseboat," etc.
Search for "?oa?" to get "goat," "croak," or "approach," etc.*
* Many PERL wildcards work as well
Select the features that you want CLEARPOND to output
Features
Neighborhood Size (count)
Mean Neighborhood Frequency
Neighbors (list of words)
Neighbor Type
Orthographic
Phonological
Neighbor Metric
Total
Substitution
Addition
Deletion
Neighbor Frequency
All Neighbors
Higher Frequency Only
Cross-Linguistic Neighbors
Dutch
French
German
Spanish
CLEARPOND automatically outputs Orthographic and Phonological Word Lengths, and Lexical Frequency (Subtlex per-million)
Optional: Supply ranges to filter your results! (format: e.g., 0-100)
Lexical Data
Lexical Frequency (Subtlex):
Word Length (Orthographic):
Word Length (Phonological):
English
Orthographic Neighborhood Size:
Orthographic Neighborhood Frequency:
Phonological Neighborhood Size:
Phonological Neighborhood Frequency:
Dutch
Orthographic Neighborhood Size:
Orthographic Neighborhood Frequency:
Phonological Neighborhood Size:
Phonological Neighborhood Frequency:
French
Orthographic Neighborhood Size:
Orthographic Neighborhood Frequency:
Phonological Neighborhood Size:
Phonological Neighborhood Frequency:
German
Orthographic Neighborhood Size:
Orthographic Neighborhood Frequency:
Phonological Neighborhood Size:
Phonological Neighborhood Frequency:
Spanish
Orthographic Neighborhood Size:
Orthographic Neighborhood Frequency:
Phonological Neighborhood Size:
Phonological Neighborhood Frequency:
CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities
Copyright Northwestern Bilingualism and Psycholinguistics Laboratory, 2012. Contact: a-shook (at) northwestern (dot) edu