IN PRESS, 2017. *JOURNAL OF WRITTEN LANGUAGE AND LITERACY.*

Orthographic and Phonological Neighborhood Databases across Multiple Languages

Viorica Marian

Northwestern University

Keywords:

Databases, orthographic neighborhoods, phonological neighborhoods, cross-linguistic, CLEARPOND, neighborhood frequency, neighborhood density, language, psycholinguistics.

Acknowledgements

Preparation of this paper was supported in part by NICHD Grant RO1HD059858. I thank the organizers and participants of the *Ninth Workshop on Written Language and Literacy* in Brighton, UK, in September 2014, under the leadership of Dr. Lynne Cahill, for the invitation to present this work and for helpful discussions. I also thank James Bartolotti for conducting the analyses on lexical decision times, Silvia Lam for organizing the references for this manuscript, Jen Krizman for comments on an earlier draft, anonymous reviewers for excellent feedback, and Anthony Shook, James Bartolotti, and Sarah Chabal for assistance in preparing the presentation slides. Correspondence regarding this paper should be addressed to Prof. Viorica Marian, 2240 Campus Drive, Evanston, IL 60208, USA. Email: v-marian@northwestern.edu.

Abstract

The increased globalization of science and technology and the growing number of bilinguals and multilinguals in the world have made research with multiple languages a mainstay for scholars who study human function and especially those who focus on language, cognition, and the brain. Such research can benefit from large-scale databases and online resources that describe and measure lexical, phonological, orthographic, and semantic information. The present paper discusses currently-available resources and underscores the need for tools that enable measurements both within and across multiple languages. A general review of language databases is followed by a targeted introduction to databases of orthographic and phonological neighborhoods. A specific focus on CLEARPOND illustrates how databases can be used to assess and compare neighborhood information across languages, to develop research materials, and to provide insight into broad questions about language. In a novel contribution using large-scale databases to answer questions about language, a closer look at neighborhood effects on lexical access reveals that not only orthographic, but also phonological neighborhoods can influence visual lexical access both within and across languages. We conclude that capitalizing upon large-scale linguistic databases can advance, refine, and accelerate scientific discoveries about the human linguistic capacity.

1. Overview of Linguistic Databases

Use of a symbolic system like language is a hallmark of human cognition. Recent advances in science and technology have provided the computing power, tools, and knowledge to study language on a scale that was not possible in previous generations. Due in part to these new capabilities in information sharing and the creation of large depositories of language knowledge, the breadth and depth of what we know about language and how it impacts the human condition is rapidly growing.

Large corpora make it possible to explore language use across a variety of contexts. Currently, a number of resources are available to those who wish to study language, and many of these resources are freely available online. For instance, CHILDES, the Child Language Data Exchange System (MacWhinney, 2000), is a large database of children's speech transcripts that is frequently used in developmental research. Similarly, the OPUS, or Open Parallel Corpus (Tiedemann & Nygaard, 2004), is a collection of multiple translated corpora from the internet (e.g., subtitles, European Parliament proceedings) across many different languages. Sites like USENET (Internet Message Board database, available through the Westbury Lab) and Wikipedia (full text of Wikipedia entries, available directly from Wikipedia) contain extensive amounts of user-generated text and provide word-banks for text analysis. These corpora and others like them have been used in the construction of many different types of databases, providing valuable information about words, such as lexical frequency, age of acquisition, familiarity, or neighborhood density.

One of the most frequently used linguistic measures collected from databases is lexical frequency. Lexical frequency refers to how often the average person might encounter a word and is available for both spoken and written language use. Frequency measures have been shown to

affect linguistic processing across many tasks, modalities, and languages, and to interact with other lexical features like neighborhood density. By looking across a variety of language databases it is possible to obtain information about spoken and written lexical frequency for words across a number of languages. For instance, the SubtLex databases provide spoken word frequencies based on subtitle corpora for Dutch (Keuleers, Brysbaert, & New, 2010), English (Brysbaert & New, 2009), French (New, Pallier, Brysbaert, Ferrand, 2004), German (Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, & Böhl, 2011), Greek (Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010), Mandarin (Cai & Brysbaert, 2010), and Spanish (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011). In addition, there are other widely available written word frequency databases, such as Google's N-Gram frequency, which compiles information from books and other written sources, with processed data available in English, Spanish, and Portuguese; the Westbury Lab's USENET and Wikipedia corpora, which are constantly growing and community generated; and CELEX (Baayen, Piepenbrock, & Gulikers, 1995), which is a well-known frequency database for Dutch, English, and German.

Lexical frequency is, of course, not the only feature of a word that can influence language comprehension and production. A number of other variables, like age of acquisition, concreteness, and familiarity play an important role in linguistic processing and information about some of these measures is also available from public databases. For example, the Center for Reading Research at Ghent University provides Age of Acquisition norms – that is, the average age at which a native speaker learns a given word – for English (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) and Dutch (Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014). Through such databases, it has been possible to understand how age of acquisition can influence lexical processing by showing, for example, that words learned first are easier to access

than words learned later in life, independent of frequency of exposure. Other important features of a word include things like how familiar a word is to a reader, whether that word can be easily pictured, and whether it represents a concrete notion, like 'cat,' or an abstract one, like 'truth.' Several databases exist for these metrics, including the MRC Psycholinguistic Database (Coltheart, 1981), the Center for Reading Research (UGENT) database (Brysbaert et al., 2014), the Hoosier Mental Lexicon (Nusbaum, Pisoni, & Davis, 1984), and B-PAL – Spanish (Davis, 2005); however, they are limited to a handful of languages.

While databases of frequency focus on features of single words, there are also databases that provide information about how words *relate* to one another. One such database is WordNet (*wordnet.princeton.edu*), which provides definitions, synonyms, and networks of word-relationships based on conceptual-semantic links. Another useful database for conceptual and lexical relations is FeatureNorms (McRae, Cree, Seidenberg, & McNorgan, 2005). Likewise, several databases attempt to capture the conceptual relationships between words based on free associations. Projects like the Small World of Words (De Deyne & Storms, www.smallworldofwords.com) are attempting to measure these relationships across multiple languages (including English, German, Dutch, French, Spanish, Japanese, Cantonese and others); however, only small parts of their data have been released. Other semantic association databases include the South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998) and the Edinburgh Associative Thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973). Finally, there are also computational databases, like the Westbury Lab High Dimensional Explorer database (HiDex; Shaoul & Westbury, 2010), which looks at the frequency with which words co-occur with one another in text as a measure of their conceptual similarity, and the Corpus of Contemporary American English (CoCA; Davies, 2009).

2. Introduction to Neighborhood Databases

Whereas semantic relationships capture conceptual similarities between words, relationships among words can also be derived from spelling or sound similarities, and encompass words that differ from one another in a single letter or sound.

An important metric that captures lexical patterns in language is lexical neighborhood. A word's neighbors are those that differ from it in either one letter (for orthographic neighborhoods) or one sound (for phonological neighborhoods). As an example, an orthographic substitution neighbor of the word PLANT can be generated by changing the last letter to form the word PLANK. Furthermore, PLANT can be changed by adding a letter, for example an E to create the word PLANET, which makes PLANET an addition neighbor to PLANT. Or, deletion neighbors can be created by subtracting a letter, for example to make PLAN. The collection of all the words that differ from a target word in one letter is called the orthographic neighborhood. Some words, like PLANT, have quite a few neighbors, so they are described as having large neighborhoods. Other words, like ORANGE, have very few neighbors and are therefore described as having small neighborhoods.

This process of adding, deleting, and substituting letters to form neighbors applies to phonology as well. For example, a substitution neighbor of DOUGH would be SHOW, an addition neighbor would be DOME, and a deletion neighbor would be OWE. A phonological neighborhood would therefore be formed by finding all of the words that differ from a target in only one sound. In English, for example, given its 26 letters, there are 456,976 possible four-letter combinations (26^4), but not all of them are possible words in English, and only 2,200 (less than half a percent) are estimated to be used based on CLEARPOND and SubtLex-us. These four

letter words have an average of 10.33 neighbors, and while certain sequences of letters appear quite frequently—such as CE, LY, others do not appear at all—such as ZW, FD (Bartolotti & Marian, 2016a).

Orthographic and phonological neighborhood information is important because neighbors and neighborhood size impact how words are processed. A simplified way to think about it is that orthographic neighbors can influence performance on visual tasks such as written word recognition (Andrews, 1992) and phonological neighbors can influence performance on auditory tasks such as spoken word recognition (Luce & Pisoni, 1998). In reality, the effect of orthography is not limited to visual word processing but also extends to auditory processing, and, similarly, the effect of phonology is not limited to auditory word processing but also extends to visual processing. In fact, the two types of neighborhoods, orthographic and phonological, can have an effect across both the visual and auditory domains. For example, during visual word recognition, large orthographic AND large phonological neighborhoods are facilitative (Grainger, Muneaux, Farioli & Ziegler, 2005; Yates, 2005; Ziegler, Muneax, & Grainger, 2003). In sum, phonological neighborhoods can and do impact performance on a task that is primarily orthographic, like reading or writing, and orthographic neighborhoods can and do impact performance on a task that is primarily phonological, like spoken comprehension or production. Given the role of phonological and orthographic neighborhoods in linguistic processing, it is necessary that this information is taken into account when conducting research on language.

Currently, there are several resources available that provide neighborhood information for English, either in the written or spoken domain. Neighborhood information derived from written corpora can be found in the MCWord Database (http://www.neuro.mcw.edu/mcword/), N-Watch, and the Washington University Speech and Hearing Lab database

(http://neighborhoodsearch.wustl.edu/Neighborhood/Neighborhome.asp). Neighborhood information derived from spoken corpora can be found in the IPHOD (http://www.iphod.com/), which provides detailed phonological neighborhood density and phonotactic probability measures.

By utilizing the information contained within these databases, it has been shown that effects of neighborhood size emerge across a variety of tasks, including lexical decision (Andrews, 1992; Luce & Pisoni, 1998; Vitevitch & Luce, 1999; Yates, Locker, & Simpson, 2004; Ziegler, Muneaux, & Grainger, 2003), reading (Tsai, Lee, Lin, Tzeng & Hung, 2006), and word learning (Frisch, Large, & Pisoni, 2000; Luce & Large, 2001; Thorn & Frankish, 2005; Roodenrys & Hinton, 2002; Storkel, Armbruster, & Hogan, 2006). For example, lexical decision tasks are completed more quickly and accurately for words with large neighborhoods, and large orthographic neighborhoods lead to faster eye movements during reading (Andrews, 1992; Luce & Pisoni, 1998; Vitevitch & Luce, 1999; Yates, Locker, & Simpson, 2004; Ziegler, Muneaux, & Grainger, 2003). Notably, even novel words show neighborhood effects (Frisch et al., 2000; Luce & Large, 2001; Thorn & Frankish, 2005; Roodenrys & Hinton, 2002; Storkel et al., 2006), with new words with dense neighborhoods easier to learn than new words with sparse neighborhoods.

However, the patterns of neighborhood effects are not always consistent *across* languages. An illustrative example is that in the case of picture naming, dense phonological neighborhoods *speed up* word production in English and Dutch (de Groot, Borgwaldt, Bos, & van den Eijnden, 2002; Marian & Blumenfeld, 2006), but *slow down* naming in Spanish (Vitevitch & Stamer, 2006). It is not clear what drives these differences, but Vitevitch and Rodriguez (2006) and Vitevitch and Stamer (2006) proposed word length and inflection (lexical

stress) as possible explanations. In Spanish, words are considered to be highly inflective, meaning that orthographic segments hold both morphologic and semantic cues. So in a language like Spanish, words that are neighbors of each other may be more related in meaning than words that are not. For example, 'niño' (little boy) and 'niña' (little girl) are not only neighbors, but also have similar meanings. In contrast, the English words 'cat' and 'can' are neighbors but are not otherwise related. This could be because, in Spanish, more of a word's neighbors may refer to related concepts (like niño and niña). It is therefore important to be are aware that neighborhoods influence processing differently across languages, especially when working with bilinguals or with cross-linguistic data.

In fact, research with speakers of more than one language must take into account not only the fact that neighborhoods vary in how they influence processing *in* different languages, but also the fact that neighborhoods exert influences *across* languages. For example, processing of the English word FACE by an English-Spanish bilingual is influenced not only by its English neighbors (such as the orthographic neighbor *lace* and the phonological neighbor *phase*), but also by its neighbors in Spanish (such as the Spanish orthographic neighbor "hace" and the Spanish phonological neighbor "seis"). But identifying and studying neighbors across languages has proven to be a challenging task, for two reasons.

First, neighborhood databases for languages other than English are scarce, and are currently limited to Basque and Spanish, with E-Hitz for Basque, and B-Pal and EsPal for Spanish (http://www.bcbl.eu/databases/espal/). Second, databases that simultaneously include neighborhood information across multiple languages are virtually non-existent. So even for languages like English and Spanish where single-language databases are available, it is still difficult to make cross-linguistic comparisons due to differences in the design of these single-

language databases, and differences in the size and quality of the data used for database construction. Therefore, unifying neighborhood information about multiple languages is a first step in being able to make cross-linguistic comparisons on this measure. WordGen, which uses comparable corpora to provide neighborhood size, frequency, and bigram information for English, Dutch, French, and German, is one such database that provides linguistic measures for multiple languages (Duyck, Desmet, Verbeke, & Brysbaert, 2004). However, WordGen is limited to within-language (not across-language) comparisons, it does not provide neighbors or neighborhood information based on phonological relationships, and is only available in a Windows format.

The need for a neighborhood database that includes multiple languages and allows for cross-linguistic comparisons served as a catalyst for developing the Cross-Linguistic Easy Access Resource for Phonological and Orthographic Neighborhood Densities, or CLEARPOND (Marian, Bartolotti, Chabal & Shook, 2012). Our goal was to create a high-quality resource that scientists who study different languages could use for bilingualism/multilingualism research and for cross-linguistic research, a resource that allows for both within-language and across-languages neighborhood analyses of phonological and orthographic neighborhoods for multiple languages.

3. The CLEARPOND Database: Cross-Linguistic Easy Access Resource for Phonological and Orthographic Neighborhood Densities

CLEARPOND includes five commonly studied languages: Dutch, English, French, German, and Spanish (see Figure 1). It calculates both orthographic and phonological neighborhood sizes for words based on phonetic transcriptions using eSpeak (a compact open

source software speech synthesizer). CLEARPOND addresses a significant gap in the field by providing for the first time cross-linguistic neighbors of words both within and between languages in an easy-to-use format. The database is now publicly and freely available as a searchable website and downloadable text files. It is designed to be useful not only to researchers working in a single language, but also to those spanning multiple languages, either through cross-linguistic or bilingual research. In CLEARPOND, neighborhood calculations take into account word frequency using the SUBTLEX databases, which are based on movie and TV subtitles and are very good at capturing frequency effects on language processing tasks.

Each word in the database includes its basic descriptors -- the word itself, its phonological transcription, its length in letters and in phonemes, and its frequency per million from the corresponding SUBTLEX frequency database. For Dutch, the SUBTLEX-NL database was used (Keuleers, et al., 2010). For English, the SUBTLEX-US database was used (Brysbaert & New, 2009). For French, the LEXIQUE (New, et al., 2004) database was used. For German, the SUBTLEX-DE database was used (Brysbaert et al., 2011). For Spanish, the SUBTLEX-ESP database was used (Cuetos, et al., 2011).

In addition to the basic descriptors, the database includes a number of other neighborhood measures, such as neighborhood size (the number of words in the list that differ from the given word by only one letter), the list of neighbors themselves, and the average frequency of those neighbors. These measures are reported both as the total number across all neighbor types, and broken down into substitution, deletion, and addition neighbors. Moreover, CLEARPOND includes two other classifications--the neighbors' modality (orthographic and phonological) and frequency threshold (higher and lower frequency). The frequency classifier provides two options, one is the total number of neighbors, and the other is restricted to

neighbors with a frequency higher than the base word; this restricted set includes the neighbors that have the largest impact on target word processing.

One of the most valuable contributions of CLEARPOND is that it allows researchers to calculate a word's neighbors across languages. For example, if one was to enter the English word FINEST into the CLEARPOND database, the results would reveal that FINEST has one neighbor in each of the five languages--*funest* (disastrous) in Dutch, *fines* in English, *fines* (fine) in French, purposes in Spanish, *findest* (find) in German--a combination of substitution, deletion, and addition neighbors. This kind of information can be used experimentally if one wants to compare words in a language that do or do not have large neighborhood sizes in another language, or as a way to control stimuli in a study on bilingual language processing.

In total, each of the twenty-seven thousand words in each of the languages in the CLEARPOND database has 240 columns of neighborhood information. There are 3 neighborhood metrics (number, frequency, and list of neighbors), and these are available for each combination of 4 neighborhood types (substitution, addition, deletion, total), 2 frequency thresholds (all words, higher-frequency words only), 2 modalities (orthographic, phonological), and 5 languages (Dutch, English, French, German, Spanish). When a user visits the CLEARPOND website, and clicks on any of the PONDS in the toolbar at the top of the page (the Dutch POND, the English POND, the French POND, the German POND, or the Spanish POND), s/he will be able to input either a single word (or nonword) or a list of words (or nonwords) and will be able to specify whether s/he wants to search within a single language or across languages. The selections provide a lot of flexibility for deciding what information is returned by the database.

4. Using Neighborhood Databases in Research

There are several valuable ways in which neighborhood databases like CLEARPOND can be used. One way to use CLEARPOND is to control for the lexical properties of a list of stimuli. Another way is to develop materials for experiments in order to answer questions about language and cognition. And perhaps the most interesting and impactful way to use CLEARPOND and other large databases is to generate and answer questions about language structure, language use, language interaction, and language change. Each of these ways is considered next.

Controlling for potential confounding variables is probably the easiest way to use the database. If a researcher has a list of words that s/he plans to use in a study, and wants to make sure the words are controlled for properties such as lexical frequency or neighborhood density, CLEARPOND will provide that information very quickly. Once at the website, the user can enter a list of words into the search box, and leave the "search by" option set to either 'Orthography' or 'Phonology' or both, depending on what needs to be controlled. One can then select Neighborhood size as the output feature so that the database will produce the neighborhood density information for the search words. Then, in the metric section, one can choose to get either the total neighbors, which is the sum of all types of neighbors, or specific neighbor types. Finally, one can query the other languages for words that are neighbors of the words entered in the target language. In the output, the results will be color-coded for each language and can be exported and saved in a text file that can easily be opened in any text editor or in spreadsheet software like Excel. This makes it very easy to save searches, and provides the data in a format that is readily available for analysis.

In addition to providing information about real words, CLEARPOND can also generate neighborhood information for non-words. Users can simply enter the non-words in the search box, as they would a word list, and CLEARPOND will calculate information about these non-words' real-word neighbors in seconds. Furthermore, users can select any additional languages when calculating non-word neighborhood information and CLEARPOND will not only produce the neighborhood information for those languages, but it will also include an alert if any of the non-words are actual words in another language. For example, a neighborhood search for the non-words 'sparf,' 'geist,' and 'wug' in English will identify 'geist' as a word in German (see Figure 2). Being able to recognize that a non-word is a real word in another language is important if, for instance, one is running a lexical decision task with bilinguals, so as to avoid the possibility that a non-word in one language is a word in another language. Note that the database can also be searched with a list that contains both words and non-words at the same time, and export them into the same file for processing.

Beyond controlling experimental stimuli for potentially confounding variables, CLEARPOND can be used to generate new stimuli for experiments. There are two primary ways to do this. For example, in an experiment investigating the effect of orthographic neighborhood size in French-English bilinguals, one may want to explore the impact that cross-linguistic neighborhoods have on reading times. Furthermore, doing this across both of bilinguals' two languages would make it possible to see if there is an asymmetric effect of neighborhoods depending upon which language people are reading in. To accomplish this goal, CLEARPOND would be used to generate four lists: two lists of English words, one where the words have many French neighbors, like 'arms', and one where the words have no French neighbors, like 'gift,' and two lists of French words, one with large English neighborhoods, like *COO,* and one with no

English neighbors, like *SUE*. This will make it possible to compare the impact of French neighborhood size on English reading times and the impact of English neighborhood size on French reading time. To generate these lists, the English POND would need to be used. The first thing to do is leave the search section blank – by doing this, the database will return all the words that meet the required criteria. If no criteria are set and the search section is blank, then it will simply return the entire database. To focus the search, parameters such as word length can be defined in the fields below the search box. For example, the word length in our hypothetical experiment could be set to 4 and the orthographic neighborhoods could be set to a range of 5 to 10 for English and 5 to 10 for French. When these selections are made, the output from CLEARPOND will return all of the words in the database that are exactly 4 letters long, have between 5 and 10 English neighbors, and have between 5 and 10 French neighbors. The next step would be to return to the English POND and set the French neighbors to zero to get the second list, English words with no French neighbors. The next step would be to go to the French POND and repeat the same process as with the English POND to obtain the data files that contain stimuli for the experiment. The entire process takes about 5 minutes, something that in the past would have taken many days of work. And because the filters for each language are stackable, the user can generate stimuli that are as precise as needed. For instance, if one wanted to restrict the lexical frequency in the output, in order to ensure equal frequency across lists, that's as easy as entering one more number during search.

Another way in which researchers can use CLEARPOND to generate new stimuli is to search the database based on elements of the words themselves, by searching with wild-cards. Suppose someone was interested in designing a study to explore how Spanish-English bilinguals read English words that contain uncommon spellings for familiar sounds. For instance, the onset

consonant cluster ‘ph’ is legal in English, but is not legal in Spanish; however, when reading English, ‘ph’ represents a sound that *is* a legal phonetic onset in both languages – the sound /f/, as in ‘fire’ or the Spanish ‘fuego’. One may wish to contrast ‘ph’ with a consonant cluster in which both the orthographic and phonetic onset is illegal in Spanish, as is the case with ‘st.’ To obtain a list of words that contain these onset clusters, the user would need to enter the consonant clusters ‘ph’ and ‘st,’ each followed by a question mark, in the search box for the English POND. CLEARPOND will then produce all of the words in the database that start with ‘ph’ and ‘st’ clusters in a single, downloadable list. And, if necessary, the list can be controlled further for lexical properties like neighborhood size or frequency.

Large lexical datasets like CLEARPOND can also be used to explore new questions about language that may be difficult or impossible to address with behavioral research. Information about average neighborhood size, or cross-linguistic neighborhoods, or word length, etc., can be used to inform us about the structures and evolution of different languages. For example, there are over a million possible five-letter combinations in English, but only a small proportion of those are actually used. If words were random combinations of letters, finding neighbors would be highly improbable. Instead, what we see with CLEARPOND is that each of its five languages has an average neighborhood size of two to three words, due to repetition of certain sequences of letters between words. CLEARPOND can also reveal how languages differ in the frequency of substitution, deletion, and addition neighbors within and across languages (Marian, Bartolotti, Chabal, & Shook, 2012). It shows that substitution neighbors are much more common than deletions or additions, especially for English. This suggests that efficient use of English graphemic space, as single letter substitutions are more likely to yield a new word compared to the other languages in the database. On the other hand, French gets a relatively large

proportion of its overall neighbors from deletions and additions. For phonological neighborhoods, the pattern is different. Dutch, German, and Spanish all have more phonological than orthographic neighborhoods. Compared to Dutch, German, and Spanish, English has nearly twice as large of a phonological neighborhood and French has over four times as many . One reason for this large difference in French is its high number of homophones. French has a very high proportion of silent letters, and as a result, many differently spelled words sound the same. For example, the French word "mer" meaning "ocean" is a phonological neighbor of four words—*ver*, *vers*, *vert*, and *verre*, all of which are homophones and are pronounced "ver," although only one of those is an orthographic neighbor. CLEARPOND reveals that over a third of the French words are homophones, compared to only two to three percent in each of the other four languages, which in itself is an interesting discovery about cross-linguistic differences in language structure. (Homophones are words that, although sounding the same in a language, carry different meanings. For example, the word *bow* is a homophone that can refer to a weapon for shooting arrows, or to a knot tied with two loops and two loose ends. Homohones are different from homographs, which are words that are spelled the same, but carry different meanings. Homophones may or may not be homographs and homographs may or may not be homophones. For instance the French word *mer* discussed above is a homophone, but not a homograph of the other four words, whereas the English word *bow* is both a homophone and a homograph across the two examples discussed. Word can have homophones, homographs, or both. For example, in addition to the homograph homophone, *bow* also has a non-homophone homograph *bow*, which refers to the bodily movement such as when greeting someone.)

CLEARPOND can also reveal interesting findings about word length. Word length strongly influences neighborhood size and is a feature that is contained within the database.

Interestingly, not only do there tend to be fewer words as length increases, but longer words are more distinct due to a larger possible inventory. (It is for this reason that autocorrect on a cell phone does a better job with longer words, which are less confusable, but frequently seems to make mistakes with short words.) While two to four letter words have the highest average neighborhood size, most neighbors in a language consist of five to eight letters, because those are the most common word lengths. CLEARPOND shows that English and French phonological word lengths are much shorter than orthographic word length, as a result of the large proportions of digraphs and silent letters in these two languages. In contrast, Spanish has nearly the same distributions of orthographic and phonological word lengths, because Spanish has a highly transparent orthography, where each letter tends to represent a single sound. German and Dutch, two highly related languages, have very similar word lengths, both favoring long words.

In using CLEARPOND to uncover differences between languages, within-language and across-language neighborhood sizes can provide insight into the distinctiveness of words in different languages. CLEARPOND shows that within-language neighbors are much more common than foreign neighbors (Marian, Bartolotti, Chabal, & Shook, 2012), which highlights the structural diversity between language pairs. English words tend to have more French neighbors, and relatively few Spanish neighbors. This is consistent with the history of English as a Germanic language that received an influx of French words through the Norman conquest in the middle ages. However, for phonological foreign neighbors, a very different picture emerges. English words have a very low proportion of foreign neighbors in any language relative to English neighbors. This is mostly because the phonological inventories of each language are distinct. Although the alphabets of these five languages remained similar (aside from the use of accent marks), their phonologies have drifted considerably from their common origins.

5. A closer look at neighborhood effects in lexical decision

To illustrate in more detail a specific way in which neighborhood databases can be used to test hypotheses about language, we used CLEARPOND to test the effect of within-language and between-language neighborhood size on visual lexical decision times. There is a dearth of research in this area, especially across languages, and existing findings are inconsistent. For example, some studies have found that lexical decision times are influenced by cross-language neighbors (van Heuven et al., 1998), while others did not find such effects (Brysbaert et al., in press).

To test the hypothesis that lexical decision times will be influenced by within-language and between-language neighborhood size, we used CLEARPOND and the Dutch Lexicon Project 2, a database of lexical decision times to 30,016 Dutch words obtained from 81 Dutch-English bilinguals. The analyses were performed on a subset of 14,972 Dutch words that were present in both the Dutch Lexicon Project 2 and CLEARPOND (excluding very low-frequency neighbors that may negatively impact the reliability and validity of such comparisons. The CLEARPOND minimum frequency cutoffs in each language are at 0.43 per million for Dutch, at 0.33 per million for English, at 0.54 per million for French, at 0.75 per million for German, and at 0.82 per million for Spanish). We used the full dataset that Brysbaert, Stevens, Mandera, and Keuleers (2016) used in their analysis of the Dutch Lexicon Project 2, including all of the predictors in their model: Word Frequency (SUBTLEX-NL and SUBTLEX-US), Word length, Age-of-acquisition (Brysbaert et al., 2014), Concreteness (Brysbaert et al., 2014), Old20 (similarity to other words, Yarkone et al., 2008; Keuleers, 2015), PoS (Part of Speech), and Word prevalence (Keuleers et al., 2015), and added Dutch and English neighborhood sizes.

The results showed that including Dutch orthographic neighborhood size improved the model, with words that have more neighbors identified more slowly, $F(1,14371) = 76.31$, $p < .001$. Adding Dutch phonological neighborhood size further improved the model, $F(2,14369) = 15.94$, $p < .001$, with phonological neighbors having a smaller effect than orthographic neighbors. There was also a negative interaction between orthographic and phonological neighbor effects, likely due to their partial overlap. That is, sometimes the same word can be both an orthographic and a phonological neighbor, but it's the number of unique neighbors that affects lexical decision. (Note that adding Dutch phonological neighborhood size alone to the base model, without Dutch orthographic neighborhood size, improved the fit, $F(1,14371) = 49.98$, $p < .001$. This difference between individual and combined model results is likely due to the overlap between orthographic and phonological neighbors.)

Next, to test neighborhood effects cross-linguistically, we added English orthographic and phonological neighborhood size information into the model. Adding English orthographic neighborhood size further improved the model, but the effect was smaller than that of Dutch, $F(1,14371) = 3.28$, $p = .07$, and, interestingly, it was in the opposite direction. English phonological neighborhood had no effect. Note that neighborhood size explains small amounts of variance after accounting for other factors. For example, the percent variance in RTs accounted for in regression analyses were as follows:

Before adding neighbors: Frequency + N_{lett} + AoA + OLD20 + Conc + Prev + PoS = 44.69%.

After adding Dutch orthographic neighbors: Frequency + N_{lett} + AoA + OLD20 + Conc + Prev + PoS + DutchOrtho = 44.98%.

After adding Dutch orthographic and phonological neighbors: Frequency + N_{lett} + AoA + OLD20 + Conc + Prev + PoS + DutchOrtho + DutchPhono = 45.10%.

After also adding English orthographic neighbors: Frequency + N_{lett} + AoA + OLD20 + Conc + Prev + PoS + DutchOrtho + DutchPhono + EnglishOrtho = 45.18%.

The small effects are not surprising. For example, in the Brysbaert et al. 2016 paper, explanatory power drops fairly steeply after frequency is accounted for. There, part of speech changed the R-squared by .0032, and in our analyses neighborhood data changed it by .0049. Although small, changes like the ones we see in the present analyses of neighborhood effects are still meaningful and are consistent with prior studies. Here, we see inhibitory effects of large Dutch neighborhoods, and facilitatory effects of large English neighborhoods. Although at first glance surprising, these results are consistent with previously-observed cross-linguistic differences. For example Grainger et al (1990) found that Dutch words with one or more higher frequency neighbors have slower RTs in a lexical decision task than Dutch words with zero neighbors, similar to our results. Van Heuven et al. (1998) tested Dutch-English bilinguals and English monolinguals on a lexical decision task for English words and found that large English neighborhood sizes were facilitatory (for monolinguals and bilinguals), and large Dutch neighborhood sizes were inhibitory (for bilinguals), similar to our results. Previous work attributed such a pattern of results to a non-target language inhibitory effect of Dutch, but here we saw the same language-specific effects when the task was performed in Dutch. Brysbaert et al (in press) also found a facilitatory effect of large English neighborhood sizes on English lexical decision tasks in monolinguals and bilinguals (although they found no effect of Dutch in their sample).

Moreover, although the participants on which the Dutch Lexicon Project was based were Dutch-English bilinguals, they also reported some knowledge of French, mostly as a result of studying it in school. To examine whether such knowledge of French influenced lexical decision

times in Dutch, we also added French neighborhood size information to the model. Adding French orthographic neighborhood size alone to the base model improved the fit, $F(1,14371) = 9.22$, $p < .01$, and adding phonological neighborhood size further improved the fit, $F(2,14369) = 4.70$, $p < .01$. This suggests that knowledge of another language, even when full bilingual status is not achieved, nevertheless impacts lexical processing in a target language.

Because Dutch and German are closely-overlapping Germanic languages, and because Spanish and French are somewhat-overlapping Romance languages, we also examined whether neighborhood size in German and Spanish influences lexical decision times in Dutch. Although we do not have information about the participants' German and Spanish exposure, we do know that they were Dutch-English bilinguals tested in the Netherlands who also studied French. Given the linguistic landscape of the Netherlands, it is likely that these participants received at least some exposure to German and Spanish, for example through availability of German and Spanish TV channels, music, and other forms of media, as well as through possible travel to German and Spanish speaking locations. Our results revealed that adding German orthographic neighborhood size alone to the base model improved the fit, $F(1,14371) = 28.03$, $p < .001$, and adding phonological neighborhood size further improved the fit, $F(2,14369) = 3.73$, $p < .05$. Adding Spanish orthographic neighborhood size alone to the base model also improved the fit, although to a smaller extent than in the case of German, $F(1,14371) = 4.38$, $p < .05$, as did adding Spanish phonological neighborhood size, $F(2,14369) = 5.61$, $p < .01$. Note however, that it is not possible to know whether the influence of German and Spanish neighborhood size on Dutch lexical decision times is due to minimal exposure to these languages (which would suggest that even minimal exposure to another language is sufficient to impact lexical processing in a native language), or whether it is a correlational artifact due to structural similarity between

languages (Dutch and German neighborhood size effects were fairly well correlated in the model, as were English and French, as well as French and Spanish, and to a lesser extent English and Spanish). It is likely that both of these variables played a role in the observed results and future research will need to disambuguate between the two variables by testing different participant samples with various language experience profiles.

A noteworthy addition from our analyses are the results concerning the phonological neighborhood effect, which previously hasn't been considered cross-linguistically in a visual lexical decision task. Here, we see that phonological neighborhood size can influence visual lexical decision times not only within, but also across languages. Finding a phonological neighborhood effect in a visual lexical decision task confirms that phonological neighborhood information is an important variable affecting orthographic processing across languages.

Together, our analyses of neighborhood effects on lexical decision show that both orthographic and phonological neighborhoods influence lexical access within and across languages. Because lexical access was in the participants' native language, these findings reinforce the argument that learning a second language impacts processing in a first language. Future research is needed to further understand how neighborhood size impacts lexical access in a second and third languages, as well as in languages other than Dutch, and to examine the contributions of language experience vis-à-vis language similarity for a more accurate understanding of language function.

6. Conclusions

These are only some of the questions about language structure, differences, similarities, interaction, and evolution that linguistic databases can provide insight into. It is clear that

neighborhood and other online databases can serve as useful tools for studying language. Using one such database, CLEARPOND, the present paper illustrates how assessing orthographic and phonological neighborhood information within and across Dutch, English, French, German, and Spanish can be used by researchers to study written and spoken processing in speakers of different or multiple languages.

We have entered a unique age in language research, created by the combination of large linguistic corpora and advanced computational power. The result is that scientists have unprecedented abilities to conduct research on a large scale and span multiple aspects within a language or across multiple languages. These capabilities to do precise, well-controlled research using large, corpus-based analyses are contributing to scientific and technological advances at a pace that is faster than ever before. It also democratizes science and discovery as it enables all of us to take advantage of the tools available online and develop new ways to expand human knowledge. The present paper discussed only a small sample of what neighborhood databases can do. To experience these capabilities first-hand, visit the links to the databases and lexica described in this paper and listed in the References section.

References

Andrews, Sally (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18(2): 234.

Baayen, R. Harald, Richard Piepenbrock & Léon Gulikers (1995). *The CELEX lexical database (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia: PA.

Bartolotti, James & Viorica Marian (2016a). Orthographic knowledge and lexical form influence vocabulary learning. *Applied Psycholinguistics.*

Bartolotti, James & Viorica Marian (2016b). Third language vocabulary learning benefits from bilingual's existing languages. *Language Learning.*

Brysbaert, Marc & Boris New (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.

Brysbaert, Marc, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte & Andrea Böhl (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.

Brysbaert, Marc, Michaël Stevens, Simon De Deyne, Wouter Voorspoels & Gert Storms (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta psychologica* 150: 80-84.

Brysbaert, Marc, Lagrou, Evelyne, & Stevens, Michael (in press). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition.*

Brysbaert, Marc, Stevens, Michael, Mandera, Pawel, & Keuleers, Emmanuel (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3), 441–458.

Cai, Qing & Marc Brysbaert (2010).SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One* 5(6): e10729.

Coltheart, Max (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology* 33(4): 497-505.

Cuetos, Fernando, Maria Glez-Nosti , Analía Barbón & Marc Brysbaert (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica* 32: 133–143.

Davies, Mark (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2): 159-190.

Davis, Colin J (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior research methods* 37(1): 65-70.

Davis, Colin J. & Manuel Perea (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods* 37 (4): 665-671.

De Deyne & Gert Storms (n.d.). *Word association study*. Retrieved from www.smallworldofwords.com

de Groot, Annette MB, Susanne Borgwaldt, Mieke Bos & Ellen van den Eijnden (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language* 47(1): 91-124.

Dimitropoulou, Maria, Jon Andoni Duñabeitia, Alberto Avilés, José Corral & Manuel Carreiras (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in psychology* 1: 218.

Duyck, Wouter, Timothy Desmet, Lieven PC Verbeke & Marc Brysbaert (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers* 36: 488–499

EsPal. Retrieved from http://www.bcbl.eu/databases/espal/

Frisch, Stefan A., Nathan R. Large & David B. Pisoni (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of memory and language* 42(4): 481-496.

Grainger, Jonathan, Mathilde Muneaux, Fernand Farioli & Johannes C. Ziegler (2005). Effects of phonological and orthographic neighbourhood density interact in visual word recognition. *The Quarterly Journal of Experimental Psychology Section A* 58(6): 981-998.

Keuleers, Emmanuel, Marc Brysbaert & Boris New (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods* 42(3): 643-650.

Keuleers, Emmanuel. (2015). VWR package for R. https://cran.r-project.org/web/packages/vwr/vwr.pdf

Kiss, George R., Christine Armstrong, Robert Milroy & James Piper (1973). An associative thesaurus of English and its computer analysis. In Adam Jack Aitken & Richard W. Bailey (eds.), *The computer and literary studies,* 153–165. Edinburgh: University Press.

Kuperman, Victor, Hans Stadthagen-Gonzalez & Marc Brysbaert (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4): 978-990.

Luce, Paul A. & David B. Pisoni (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing* 19(1): 1.

Luce, Paul A. & Nathan R. Large (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes* 16: 565-581.

MacWhinney, Brian (2000). *The CHILDES Project: Tools for Analyzing Talk.* New Jersey: Lawrence Erlbaum Associates.

Marian, Viorica & Henrike Blumenfeld (2006). Phonological neighborhood density guides lexical access in native and non-native language production. *Journal of Social and Ecological Boundaries* 2: 3-35.

Marian, Viorica, James Bartolotti, Sarah Chabal & Anthony Shook (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one* 7(8): e43230.

McRae, Ken, George S. Cree, Mark S. Seidenberg & Chris McNorgan (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods* 37(4): 547-559.

MCWord. Retrieved from http://www.neuro.mcw.edu/mcword/

Nelson, Douglas L., Cathy L. McEvoy & Thomas A. Schreiber (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36(3): 402-407.

New, Boris, Christophe Pallier, Marc Brysbaert & Ludovic Ferrand (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers* 36(3): 516-524.

Nusbaum, Howard C., David B. Pisoni & Christopher K. Davis (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words.* (Progress Report No. 10; pp. 357–376). Bloomington, IN: Speech Research Laboratory, Indiana University.

Roodenrys, Steven & Melinda Hinton (2002). Sublexical or lexical effects on serial recall of nonwords?. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28(1): 29.

Shaoul, Cyrus & Chris Westbury (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods* 42(2): 393-413.

Storkel, Holly L., Jonna Armbruster & Tiffany P. Hogan (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research* 49 (6): 1175-1192.

The Irvine Phonotactic Online Dictionary (IPhOD). Retrieved from http://www.iphod.com/

Thorn, Annabel SC & Clive R. Frankish (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(4): 729.

Tiedemann, Jörg, and Lars Nygaard (2004). *The OPUS Corpus-Parallel and Free*. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.

Tsai, Jie-Li, Chia-Ying Lee, Ying-Chun Lin, Ovid JL Tzeng & Daisy L. Hung (2006). Neighborhood size effects of Chinese words in lexical decision and reading. *Language and Linguistics* 7(3): 659-675.

Van Heuven, Walter, Dijkstra, Ton, & Grainger, Jonathan (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language, 39*(3), 458-483.

Vitevitch, Michael S. & Eva Rodríguez (2006). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders* 3 (1): 64-73.

Vitevitch, Michael S. & Melissa K. Stamer (2006). The curious case of competition in Spanish speech production. *Language and cognitive processes* 21(6): 760-770.

Vitevitch, Michael S. & Paul A. Luce (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40(3): 374-408.

Washington University Speech and Hearing Lab database. Retrieved from http://www.neuro.mcw.edu/mcword/

Westbury Lab. Retrieved www.psych.ualberta.ca/~westburylab/

WikiPedia. Retrieved from https://www.wikipedia.org/

WordNet. Retrieved from *wordnet.princeton.edu*

Yates, Mark (2005).Phonological neighbors speed visual word processing: evidence from multiple tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(6): 1385.

Yates, Mark, Lawrence Locker & Greg B. Simpson (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review* 11(3): 452-457.

Ziegler, Johannes C., Mathilde Muneaux & Jonathan Grainger (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language* 48(4): 779-793.

Figure 1.

Image of the CLEARPOND database.

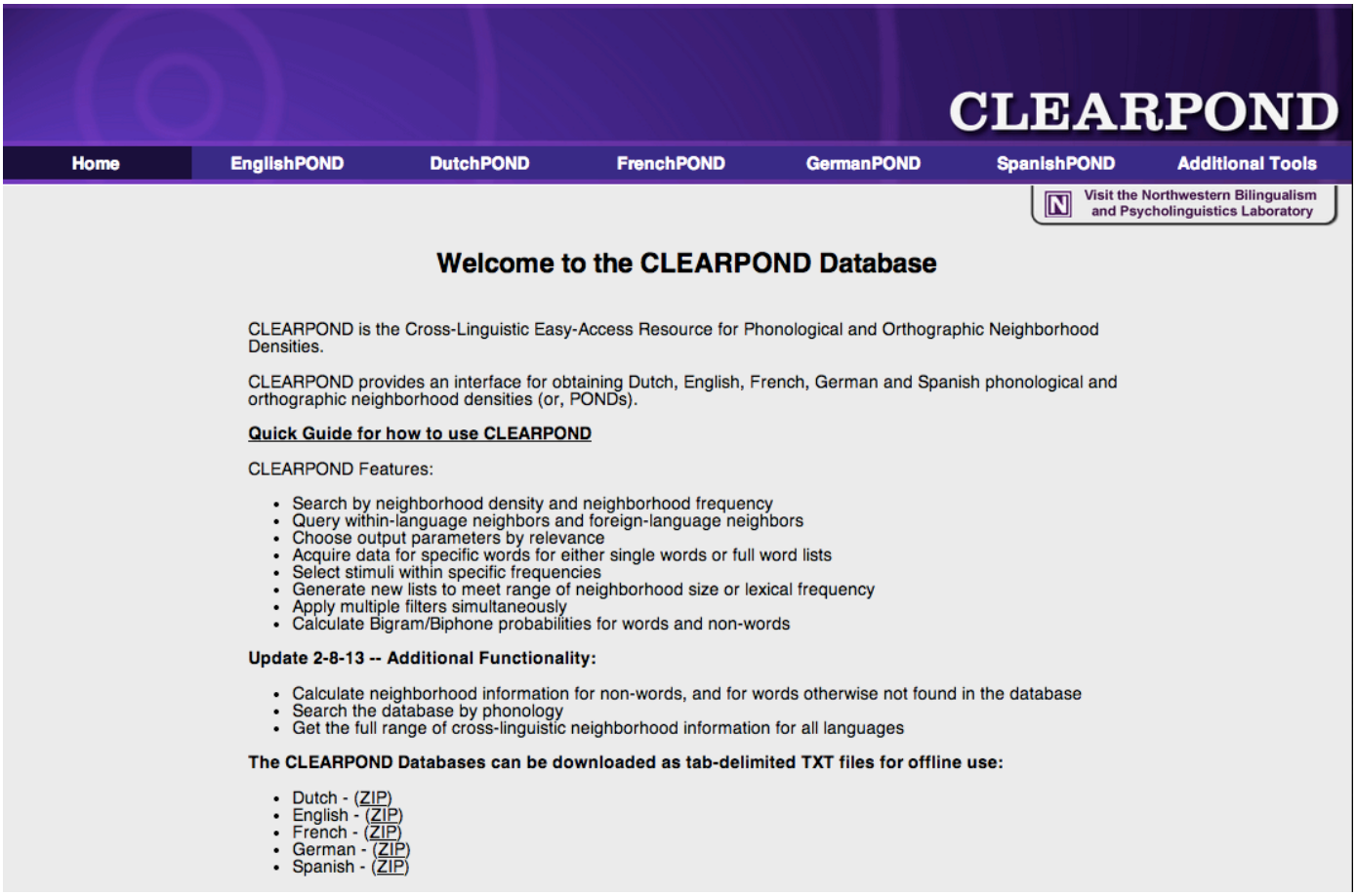
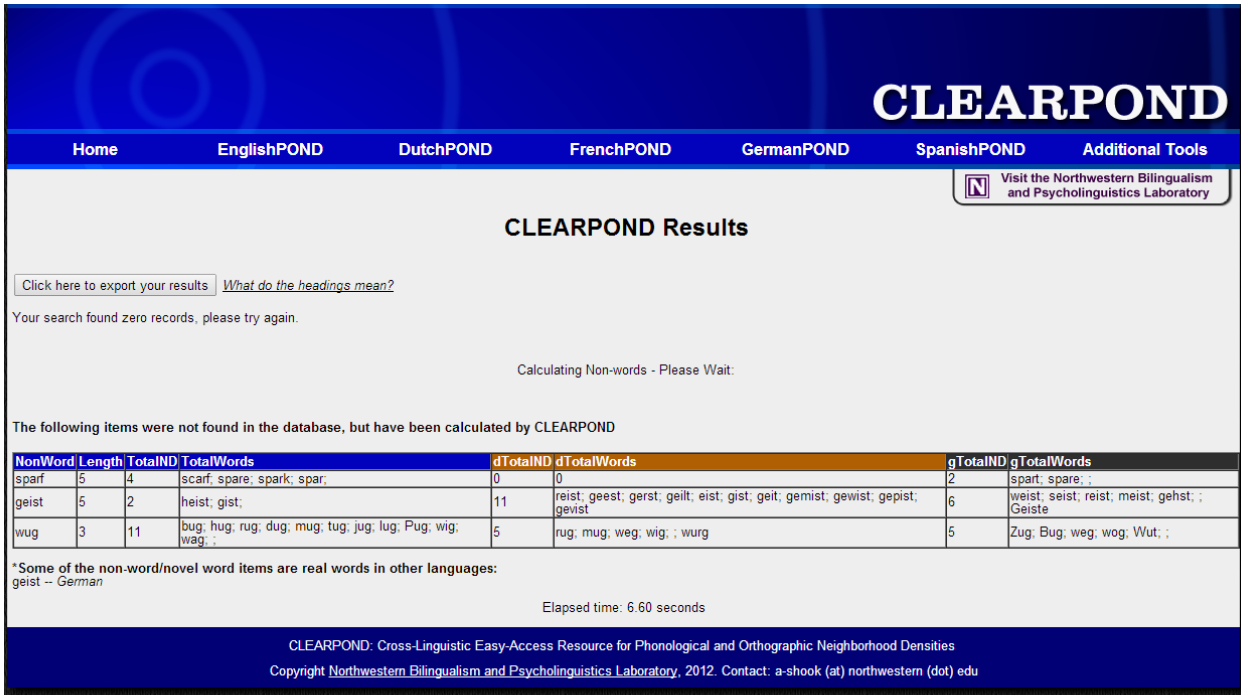


Figure 2.

Example of output provided by the CLEARPOND database when a search was performed for neighbors of three English nonwords within English and across English and German.



CLEARPOND

Home | EnglishPOND | DutchPOND | FrenchPOND | GermanPOND | SpanishPOND | Additional Tools

Visit the Northwestern Bilingualism and Psycholinguistics Laboratory

CLEARPOND Results

Click here to export your results | *What do the headings mean?*

Your search found zero records, please try again.

Calculating Non-words - Please Wait:

The following items were not found in the database, but have been calculated by CLEARPOND

NonWord	Length	TotalND	TotalWords	dTotalND	dTotalWords	gTotalND	gTotalWords
sparf	5	4	scarf; spare; spark; spar;	0	0	2	spart; spare; ;
geist	5	2	heist; gist;	11	reist; geest; gerst; geilt; eist; gist; geit; gemist; gewist; gepist; gevist	6	weist; seist; reist; meist; gehst; ; Geiste
wug	3	11	bug; hug; rug; dug; mug; tug; jug; lug; Pug; wig; wag; ;	5	rug; mug; weg; wig; ; wurg	5	Zug; Bug; weg; wog; Wut; ;

***Some of the non-word/novel word items are real words in other languages:**
geist -- *German*

Elapsed time: 6.60 seconds

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