Chapter 10 : Corpus Linguistics

Marc Brysbaert, Paweł Mandera, Emmanuel Keuleers

Ghent University

Keywords: corpus linguistics, word frequency measures, semantic vectors, big data, megastudies, crowdsourcing, ratings of word features, stimulus matching

Chapter to be published in De Groot A.M.B. & Hagoort, P. (2016). Research Methods in Psycholinguistics. London: Wiley.

Abstract

Corpus linguistics refers to the study of language through the empirical analysis of large databases of naturally occurring language, called corpora. Psycholinguists are mostly familiar with corpus linguistics because the word frequency norms they use come from corpus linguistics. The frequency norms are more informative if they include information about the part-of-speech roles of the words (e.g., the word “dance” used as a verb or a noun). This requires the syntactic parsing of the corpus, which is currently done automatically. An exciting new development is the calculation of semantic vectors on the basis of word co-occurrences. In this analysis, the meaning of a target word is derived by taking into account the words surrounding the target word. This makes it possible to calculate the semantic similarity between two target words. The measures provided by corpus linguistics are the most powerful when they can be combined with processing times for large numbers of words (obtained in megastudies) and subjective ratings for many words (obtained via crowdsourcing studies). Examples are given.

Introduction

Corpus linguistics refers to the study of language through the empirical analysis of large databases of naturally occurring language, called corpora (singular form: corpus). In linguistics, corpus linguistics for a long time was the rival of approaches that predominantly valued the theoretical insights and acceptability intuitions of individual linguists. In recent years, signs of collaboration and cross-fertilization are observed (Gries, 2010), partly because the tools used in corpus linguistics have become more user-friendly. Everyone looking up the use of a particular phrase on an internet search engine is essentially doing corpus linguistics, searching a large collection of webpages for the presence of a particular word or word co-occurrence. At the same time, ideas from theorists are important for corpus linguists, as corpus searches are particularly informative when they address specific, theory driven predictions.

Psycholinguists are mostly familiar with corpus linguistics because of the word frequency measures they use. It is well-known that high-frequency words are processed more efficiently than low-frequency words. The frequency norms, on which the selection of stimulus materials is based, come from corpus linguistics. In particular, the compilation of a balanced, one million word corpus by Kucera and Francis (1967) and the word counts based on that corpus have had a tremendous influence on word recognition research in English up to the present day. Corpus analysis also has had an influence on sentence parsing research, first to find out which constructions were attested and which not, then to find out the relative frequencies of various constructions, and now increasingly to train computational models of sentence parsing. Another exciting use of corpus analysis is the calculation of semantic similarity measures on the basis of word co-occurrences.

Assumptions and Rationale

The underlying assumptions of corpus linguistics differ slightly between studies depending on whether a researcher is interested in language production or language perception. For language production researchers, the corpus is the output to be analyzed and the ideal is to have the largest possible sample of spontaneously generated contents. These can be written texts, but most of the time will consist of spoken discourse, as there are more researchers interested in speech production than in writing, and because written texts are often edited and polished before publication (although there are exceptions, such as television programs that are subtitled online or chat interactions). The rationale behind the approach is that the corpus forms a representative sample of language produced and, therefore, can be analyzed to reveal the processes underlying language production. Typical examples of such studies are the analysis of speech errors (e.g., saying “dye a beggar” instead of “buy a dagger”; Fromkin, 1973) or the investigation of acoustic reductions in speech (Ernestus, Baayen, & Schreuder, 2002).

The main assumption made by word perception researchers is that the language corpus is representative for the type of language people have been exposed to in their lives. The corpus can then be used to count the frequencies of various words, phrases, and syntactic constructions encountered by people. This has been the basis of all research on word frequency (Brysbaert & New, 2009; Monsell, Doyle, & Haggard, 1989). It has also been the basis of all research investigating whether people are more likely to use the most frequent analysis when confronted with a syntactic ambiguity (Reali & Christiansen, 2007).

A criticism raised against the rationale behind using frequency measures in perception research is that a correlation between frequency of production and ease/preference of use need not be interpreted as evidence for the claim that exposure drives perception. It can be defended that exposure frequency does not affect interpretation directly, but that both are

the outcome of a third variable influencing both production and perception. For instance, it has been argued that differences in structural complexity and working memory demands drive both syntactic production and perception: One is likely to produce the structure with the least demands and one tends to prefer the disambiguation with the simplest structure. Similarly, with respect to the word frequency effect in speech production, Hayes (1988) wondered whether the observation that spoken discourse contains fewer low-frequency words than written texts could be due to people avoiding the use of low-frequency words in spoken discourse in order to preserve the fluency of their speech. According to Hayes the difficulty of producing a word determines the frequency of occurrence (and not the other way around). It is good to keep these objections in mind: A correlation between production and perception need not mean that perception is directly affected by frequency differences in the language exposed to, as assumed by experience-based models of language processing. On a more positive note, the correlation between perception and corpus data can be used to predict one from the other, independent of the underlying causality structure.

Apparatus and Tools

The apparatus for corpus linguistics is becoming simple, as a result of the growing power of computers. Most desktop and laptop computers nowadays can do the analyses that required supercomputers only a few decades ago. The most likely impediment to applying corpus linguistics is the computer programming skills required. Given that corpora currently contain billions of words/sentences, one needs automated algorithms to process the data. Indeed, there is a big overlap between corpus linguistics and natural language processing (NLP) research in departments of computer sciences, where one tries to improve the verbal intelligence of computers by making them digest large corpora of information (usually texts, although the first uses of pictorial materials have been reported). Increasingly, libraries of algorithms and

software packages become available, making it possible to run programs without requiring in-depth knowledge of the underlying operations, just like statistical packages make it possible to run complicated analyses without being familiar with matrix algebra (what Schütz, 1962, called the use of recipe knowledge). A few of the packages are mentioned at the end of the chapter. However, because the packages rapidly change and are language-dependent, our list is likely to be outdated soon and it is better to do an internet search. Two programming languages that are popular at the moment are R and Python.

Depending on the information one needs, it is possible to do direct searches in a corpus. This will be the case when one is interested in the occurrence of certain words or word sequences. In many cases, however, one will want to have more information than can be derived from a surface analysis, for instance when one is interested in syntactic structures or in part-of-speech information related to the words. For such questions, it is important to have access to a corpus that has been parsed and tagged. Parsing refers to the decomposition of sentences into their grammatical constituents, which are put into a tree diagram indicating the syntactic relationships between the constituents. Tagging involves the assignment of part-of-speech (PoS) information to the words, which includes the assignment of the right lemma (base form) to inflected words. A number of small corpora have been parsed and tagged manually (the most famous arguably is the Penn Treebank). Most of the time, however, this is done automatically now, even though the output is not yet 100% error-free. Software packages often used in English include CLAWS (http://ucrel.lancs.ac.uk/) and the Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml).

Occasionally (and way too infrequently) psycholinguists can profit from derived data made available by computer linguists or NLP scientists. As indicated above, the best known example is the availability of word frequency lists. These lists consist of word types, the number of times they have been observed in the corpus, the syntactic roles (parts-of-speech)

of the word, and the lemmas associated with these parts-of-speech (see below). This information can often be reduced to a single file for a spreadsheet or a statistical program, or made available through a website. An interesting addition in recent years is the collection of frequencies of word sequences (called word Ngrams). These consist of word bigrams (frequencies of word pairs), word trigrams (sequences of three words), and so on. They were first made available by Google (https://books.google.com/ngrams). Another interesting website for English word Ngrams is the Corpus of Contemporary American English (http://corpus.byu.edu/coca/).

Nature of Stimuli and Data

Raw data vs. derived data

The nature of the stimuli depends on whether you make use of the corpus itself or of derived data. If you want to work with a corpus yourself, you obviously must have access to it. This will consist of text, sometimes enriched with additional information such as part of speech associated with words or parse structure of the sentences included in the corpus. (Spoken materials are usually transcribed, because it is not yet possible to do corpus-wide analyses on speech signals.)

A major limitation of corpora is that most of them are subject to copyright restrictions, because the materials were produced by other people, who did not transfer copyright to the corpus builders (this is often impossible given the sheer number of people and organizations involved). Because of possible copyright infringements, researchers are very hesitant to share their corpora with colleagues, meaning that many corpora must be built anew by research groups, hindering the accumulation of information and the replication of findings.

The situation is much better for derived data, as these data usually are free for research purposes and are easier to handle. Because the derived data do not harm the authors'

commercial rights, they do not violate intellectual property and fall under the rules of "fair use of a copyrighted work". In their simplest form, the derived data are available as a spreadsheet (e.g., Excel) and can be used by anyone with basic computer skills. Occasionally, the list is too long and then you need access to (slightly) more advanced software.

Language corpora need not be limited to spoken and written words. They can also consist of gestures, either to replace speech (in mute or deaf participants) or to accompany speech.

Word frequency data

The most frequently used measure derived from corpus linguistics is word frequency. Table 1 shows an excerpt from the SUBTLEX-US database (Brysbaert, New, & Keuleers, 2012), which contains word frequencies based on an analysis of a corpus of film subtitles including 51 million words from 9,388 films. It describes the information for the word "appalled". The first line shows that this word was observed 59 times in the corpus. The second line indicates that it was observed in 53 films (a variable called "contextual diversity"). The third and the fourth line provide standardized frequency measures: Frequency per million words (59/51 = 1.16) and the Zipf-value, which is a standardized logarithmic value (log10((59+1)/51)+3 = 3.07). The Zipf-value is a better measure than frequency per million, because it takes into account the facts that the word frequency effect is a logarithmic function and that more than half of the words have a frequency of less than one per million words. The value ranges from 1 to 7, with low-frequency words covering the range of 1-3, and high frequency words covering the range of 4-7. For more information, see van Heuven, Mandera, Keuleers, and Brysbaert (2014). The next lines of Table 1 indicate that "appalled" is used as an adjective (49 times) and as a verb form (10 times). So, the dominant lemma of "appalled" is "appalled" (used as an adjective); the other lemma is the verb "appall".

Table 1. Excerpt from the SUBTLEX-US database for the word “appalled”.

Word	appalled
FREQcount	59
CDcount	53
SUBTLEX pm	1.16
Zipf-value	3.07
Dom_PoS_SUBTLEX	Adjective
Freq_dom_PoS_SUBTLEX	49
Percentage_dom_PoS	0.83
All_PoS_SUBTLEX	Adjective.Verb
All_freqs_SUBTLEX	49.10
Dom_Lemma_SUBTLEX	appalled
All_Lemma_SUBTLEX	appalled.appall

Because word frequencies are so easy to calculate nowadays, it is important to make sure you use a good frequency measure (see the next section as well). Important variables to consider are (1) the size of the corpus, (2) the language register captured by the corpus, and (3) the quality of the analyses done.

As for the size of the corpus, good frequency measures require corpora of some 20-50 million words. This is because a large part of the word frequency effect is situated at frequencies lower than one per million words (Keuleers, Diependaele, & Brysbaert, 2010). These are the Zipf-values between 1 and 3. If the corpus is too small, it is impossible to measure these frequencies properly. Larger corpora are required when in addition one wants information about part-of-speech or word Ngrams.

At the same time, it is not true that large corpora are always better than small corpora, the reason being that large corpora often tap into language registers few participants in psychology experiments (typically undergraduate students) are familiar with. Such corpora are, for instance, encyclopedias. Wikipedia is a very popular source in NLP research, because it contains nearly 2 billion words, is freely available, and exists for many languages. However, it is not the type of language undergraduates read a lot. The same is true for Google

books, which is another multibillion word corpus, covering millions of fiction and non-fiction books, but again unlikely to be read by undergraduates. When the quality of word frequency measures is tested, substantially better results are obtained when the corpus consists of film subtitles (Brysbaert, Keuleers, & New, 2011), tweets and blogs (Gimenes & New, in press), or facebook messages (Herdağdelen & Marelli, in press), as discussed in the next section.

Finally, the quality of the word frequency measure also depends on the quality of the analysis done. Several factors are involved. One of them is the multiplication of sources. Because electronic materials are easy to copy, most corpora contain multiple instances of the same information (e.g., subtitles for the same film in a corpus of subtitles). It is important to detect and delete such duplications. The same is true for interchanges where previous messages are copied in the replies. Often some checks of the text quality must be done as well, to make sure that the language is the one intended and of an acceptable level. Another issue is that files often contain meta-information related to the source, which must be discarded as well. For instance, files with film subtitles usually include information about the film, the people who made the subtitles, and so on. This information must be excluded. Lastly, if one is interested in part-of-speech information, it is important to use a parser of good quality.

The following are interesting sources across a number of languages. The first are the so-called SUBTLEX frequencies, based on film subtitles and available for Chinese, Dutch, English, French, German, Greek, Polish, Portuguese, and Spanish (for more information, see http://crr.ugent.be/programs-data/subtitle-frequencies). Another interesting source comes from tweets and blogs. Gimenes and New (in press) provide them for 66 languages. Some databases are geared towards children. The best known is the CHILDES database, available for several languages (http://childes.psy.cmu.edu/) and discussed extensively in Chapter 3.

Semantic vectors

Whereas corpus studies traditionally were geared towards word frequency data and syntactic analyses, an exciting development in the past two decades is the calculation of semantic information on the basis of word co-occurrences. This approach, which is based on the idea that words with similar meanings tend to occur in similar contexts (Harris, 1954), was introduced to psychology in two classic papers by Lund and Burgess (1996) and Landauer and Dumais (1997). The authors operationalized the semantic similarity between words by observing the joint occurrence of the words in contexts. For Lund and Burgess, the context was a small moving window (up to 10 words) sliding through the corpus. For Landauer and Dumais, the context was a short article.

Lund and Burgess compiled a corpus of 160 million words from internet news groups. In their analysis, they included all words appearing at least 50 times within the corpus. This resulted in a total of 70 thousand words and a co-occurrence matrix of 70,000 x 70,000 entries. Each cell of the matrix included the number of times the words were present together in the sliding window. On the basis of this matrix, each word had a semantic vector consisting of 70,000 numbers. By comparing the semantic vectors, the semantic similarity between words could be calculated: Words that co-occurred in the same contexts had very similar semantic vectors; words that rarely co-occurred in the same context had different semantic vectors. Lund and Burgess observed that the similarity vectors made clear distinctions between words from the categories animals, body parts, and geographical locations. The vectors of the words within these categories were much more similar than those between the categories. The authors also showed that the semantic similarities were larger between targets and related primes from a previously published semantic priming experiment than between targets and unrelated control primes. Lund and Burgess called their approach the *hyperspace analogue to language* (HAL; see also Chapter 9)).

Landauer and Dumais (1997) started from the same basic principles but applied a slightly different procedure. First, they used a corpus consisting of a 4.6 million word encyclopedia for young students, which included 30,473 entries (in later implementations the authors worked with a larger corpus of schoolbooks to better approach the learning process in children). From each entry the authors took a text sample with a maximum of 2,000 characters (about 151 words). The encyclopedia entries formed one dimension of the matrix; the other dimension consisted of the 60,768 words they were interested in. The cells in the matrix contained the frequency with which a particular word appeared in a particular text sample. Next, the authors applied a dimensionality reduction to the matrix (called *singular value decomposition*), which reduced the 30,473 entries to 300 dimensions. Again the values of the words on each of these 300 dimensions were used as a vector to calculate the similarity to the other words. To test the usefulness of the semantic vectors, Landauer and Dumais used them to solve a vocabulary test with multiple choice answer alternatives (taken from the synonym portion of the Tests of English as a Foreign Language – TOEFL). The test consisted of 80 items with four alternatives to choose from. An item was correctly solved when the semantic distance calculated between the target and the correct alternative was smaller than the distances with the other three alternatives. This was the case for 64% of the items, which agreed with the score obtained by a large sample of applicants to US colleges from non-English speaking countries. Landauer and Dumais called their approach *latent semantic analysis* (LSA; see also Chapter 9).

From a practical point of view, an important difference between Lund and Burgess (1996) and Landauer and Dumais (1997) was that the latter not only published their paper, but also developed a website (http://lsa.colorado.edu/) on which visitors could calculate the LSA similarities between words. This website informs you, for instance, that the semantic similarity between apple and pear is .29, whereas the similarity between apple and tear is .18.

The site also informs you that other words are closer neighbors to apple. Some of these are in descending order: cherry (.43), peel (.42), and tree (.40). Surprisingly, the list also includes chicle (.41), nonalphabetic (.40), uppercase (.39), and chapman (.38), showing that further improvements to the measure are warranted. Because of the availability of the user-friendly interface with derived measures, LSA has had much more impact on psycholinguistic research than HAL. Indeed, one regularly comes across semantic-priming experiments in which LSA values were compared or matched across conditions.

In the years since the publications of Lund and Burgess (1996) and Landauer and Dumais (1997), researchers have attempted to improve the performance of the procedures. Several approaches were taken. First, researchers made use of larger corpora. Second, they tried to optimize the transformation steps applied to the raw context count matrices and searched for the best possible parameter sets. One of the testing standards was the TOEFL test used by Landauer and Dumais. Gradually, the number of correctly solved items rose until Bullinaria and Levy (2012) reached 100% correct test performance. This was achieved by using a corpus of over 2 billion words crawled from the web (including Wikipedia pages), a HAL based approach with a window size of one word to the left and one word to the right of the target word, a cosine semantic similarity index, and by weighting the vector components. Lemmatizing a text before running the analysis (i.e., replacing all inflected forms by lemmas) did not improve the performance of the models if the corpus was big enough.

In addition to improving well-established models, completely new approaches have been proposed. One was that researchers started to use a connectionist network rather than a count matrix (Mikolov, Chen, Corrado, & Dean, 2013). In these models, word co-occurrences are no longer explicitly counted and reduced to principal components. Instead, all target words are represented as input and output nodes in a three-layer connectionist network. The context words are used as predictors in the input layer and the target word is the one that must

be activated in the output layer. The input and output layers are connected via a hidden layer of a few hundred units. The weights between the nodes are adapted to optimize the performance of the network and the final weights are used to form the semantic vectors (see Chapter 9 for details about connectionist models). Several studies have confirmed that this approach usually yields better and more robust performance than the traditional distributional models, such as HAL or LSA (Baroni, Dinu, & Kruszewski, 2014; Mandera, Keuleers, & Brysbaert, 2017; but see Levy, Goldberg, & Dagan, 2015 for an alternative view). In addition, it has been shown that the proposed connectionist models can be mathematically equivalent to a certain type of the traditional models (Levy & Goldberg, 2014). At the same time, it has been suggested that better performance on the TOEFL may not be the best indicator of human performance, because optimal performance on the TOEFL test requires encyclopedic input, whereas human semantic priming data are better predicted by semantic vectors based on everyday language such as found in film subtitles (Mandera et al., 2017).

Unfortunately, the access to the information and skills needed to independently train and use the state-of-the-art semantic vectors make them out of reach to many psycholinguistic researchers. The corpora on which the new measures were calculated cannot be made freely available due to copyright restrictions, and running the algorithms requires expert knowledge (not to mention computer time). As a result, psycholinguists had little option but to continue working with the easily available but outdated LSA measures. To solve this problem, we have written a shell that can be downloaded and calculates the semantic distances between words based on the latest developments (http://crr.ugent.be/snaut/). At the moment, the shell calculates semantic distance values for English and Dutch. Other languages are likely to follow.

Collecting the Data

Most of the time, a corpus will be downloaded from the internet. Indeed, the massive availability of language in digital form has been the driving force behind corpus linguistics. Researchers have a tendency to go for the materials that are easiest to reach. As indicated above, a lot of corpora contain the Wikipedia webpages (https://www.wikipedia.org/), as they can be downloaded easily. This is a good corpus for encyclopedic knowledge, but is less suited as a proxy for the typical speech or text people are exposed to. Some other popular text corpora are based on web crawlers that browse the World Wide Web and download the contents of various sites. These corpora contain a wide variety of sources (which is good), but usually require considerable cleaning (duplicates, pages in other languages, pages with repetitions of the same information, etc.). Finally, some corpora can be obtained from previous research (but see the copyright issues above). The advantage here is that much of the cleaning work has been done already.

The size required for a good corpus depends on its use. If the goal is to have frequencies of single words, then a corpus of some 20-50 million words is enough (Brysbaert & New, 2009). If one in addition wants reliable part-of-speech information about low-frequency words, a larger corpus is indicated. Larger sizes are also needed if the researcher wants information about word co-occurrences, as these are by definition lower in frequency. At the same time, it is good to keep in mind that an undergraduate student (the typical participant in psycholinguistic experiments) is unlikely to have come across more than 2 billion words in their life (Brysbaert, Stevens, Mandera, & Keuleers, 2016a). So, corpora larger than this size are less representative as well.

Next to size, the language register of the corpus is of critical importance, certainly if one wants to predict performance in psycholinguistic experiments. In general, measures based on the type of language participants have been exposed to are more valid than measures based on scientific or non-fiction sources. As indicated above, particularly useful sources are film

subtitles and social media messages. Also school books are a good source, arguably because undergraduates spent a good part of their lives reading and studying them. Books from primary school have an extra advantage because they tap into the language first acquired, which seems to have a stronger influence on language processing than words acquired later (Brysbaert & Ellis, in press). A special case concerns research with participants of old age, as these have been less exposed to internet language and language from recent years. Several studies report that for these participants, corpora of some time ago may be more representative (for references, see Brysbaert & Ellis, in press).

The register of the corpus is particularly relevant when one wants to compare the processing of various types of words. One such question is whether emotional words (associated with positive and negative feelings) are recognized faster than neutral words. To answer this question, one must be sure that the frequencies of the various words are estimated correctly (Kuperman, Estes, Brysbaert, & Warriner, 2014). For instance, if the word frequency estimates are based on a non-fiction corpus, the frequency of the emotional words will be underestimated (as non-fiction texts rarely deal with emotion-laden situations) and it will look as if emotional words are processed faster than expected on the basis of their "frequency". Alternatively, if the corpus is based on song lyrics, it might seem like emotional words are processed more slowly than expected on the basis of their "frequency".

An Exemplary Application

There are two ways to show the utility of the various measures provided by computational linguistics: either by setting up a new study that addresses a specific theoretical question or by reanalyzing an old study. We take the latter approach and consider the stimuli used in a randomly chosen semantic priming experiment (de Mornay Davies, 1998, Experiment 1). The

experiment was based on 20 target words preceded by semantically related and unrelated primes. These are shown in the first three columns of Table 2.

Table 2. Stimuli used in a semantic priming experiment by de Mornay Davies (1998). The first three columns show the stimuli (target, related prime, unrelated prime). The fourth column gives the SUBTLEX-UK frequency of the target word (expressed in Zipf-values) and the fifth column gives the dominant part-of-speech of the word.

TARGET	**RELATED**	**UNRELATED**	**$Zipf_{target}$**	**$DomPoS_{target}$**
bird	wing	shirt	4.85	noun
bottle	glass	claim	4.65	noun
boy	girl	land	5.28	noun
chase	run	town	4.31	verb
cup	plate	pitch	5.09	name
drop	fall	club	4.90	verb
fast	slow	goal	5.09	adverb
gammon	bacon	spade	2.85	noun
glove	hand	think	3.81	noun
house	home	small	5.83	noun
lance	sword	canoe	3.74	name
light	dark	view	5.28	noun
lock	key	add	4.42	noun
mail	letter	effort	4.63	noun
moon	sun	shot	4.74	noun
string	rope	clue	4.25	noun
tail	feather	parent	4.45	noun
wash	clean	sweet	4.54	verb
wig	hair	food	3.82	noun
wire	cable	tiger	4.29	noun

The first thing we want to know about these stimuli is their word frequency. As the experiment was run in the United Kingdom, we want frequencies for British English. A good source for these are the SUBTLEX-UK frequencies (Van Heuven et al., 2014). They can be found at the website http://crr.ugent.be/archives/1423. The 4th column of table 2 shows the outcome for the target words. The mean Zipf value is 4.54 (SD = .67), which is rather high (similar to a frequency of 28 per million words). It is further noteworthy that the targets consist of a combination of nouns, verbs and adverbs, with two words that are primarily used

as proper nouns (Cup, Lance). These are stimuli we may want to avoid in a good experiment. A similar analysis of the related primes shows that their average Zipf frequency is 4.84 (SD = .50), that they include one word mostly used as a proper noun (Cable) and four words mostly used as an adjective (clean, dark, key, slow) in addition to nouns. The frequency of the unrelated primes is 4.85 (SD = .67), well matched to the related primes. They include two verbs (claim, think) and two adjectives (small, sweet), in addition to 16 nouns.

It is furthermore interesting to see how much the related and the unrelated primes differ in semantic distance. We use the semantic vectors of Mandera et al. (2017). The semantic distance is .50 (SD = .12) between the targets and the related primes (on a scale going from 0 – fully related – to 1 – fully unrelated). The distance between the targets and the unrelated primes is .84 (SD = .09), which is substantially higher.

In addition to the above measures, we could also check whether the stimuli are well matched on other variables known to influence visual word recognition, such as word length, age-of-acquisition, and orthographic/phonological similarity to other words. For English, information about the similarity to other words can be looked up in Balota et al. (2007; http://elexicon.wustl.edu/) or calculated with the vwr package (Keuleers, 2015). Information about age-of-acquisition can be found in Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012; http://crr.ugent.be/archives/806). Applied to the data of Table 2, the orthographic similarity to other words, as measured with OLD20 in Balota et al. (2007), is 1.40 (SD = .26; the word ‘gammon’ is not in the database) for the target words, 1.49 (SD = .29) for the related primes, and 1.71 (SD = .29) for the unrelated primes. The deviation of the last value indicates that better primes could have been chosen in the unrelated condition. The age-of-acquisition values are 4.96 (SD = 3.09) for the targets, 4.27 (SD = 1.66) for the related primes, and 5.58 (SD = 1.64) for the unrelated primes, again suggesting that a better matching of the prime stimuli is possible.

In summary, the stimuli used by de Mornay Davies (1998, Experiment 1) were not bad, but they can be further improved, so that they all consist of nouns, and are fully matched on variables such as orthographic similarity (OLD20) and age-of-acquisition. Having access to databases such as those just mentioned allows us to run better controlled experiments in psycholinguistics. Such information can also be used in regression analyses based on processing times for thousands of words, to find out the relative impact of the various variables (Keuleers & Balota, 2015; Brysbaert, Stevens, Mandera, & Keuleers, 2016b).

Limitations and Opportunities for Validation

Corpus linguistics provides psycholinguists with valuable tools to investigate language processing. Research on word processing would be impossible without access to word frequency information, morphological information, and word similarity indices, all based on corpus analyses.

A new extension that is currently tried out is to see how well specific word features can be calculated on the basis of semantic vectors. For instance, it seems reasonable to derive the emotional value of a word from the emotional values of its semantically close words. If one knows that the word ‘beautiful’ has a positive affect, one can be pretty sure that the same will be true for all its synonyms, such as ‘lovely’, ‘attractive’, ‘good-looking’, ‘gorgeous’, ‘stunning’, ‘striking’, and ‘handsome’ So, by using a limited number of seed words and semantic similarity vectors, it may be possible to estimate the emotional value of all words in a language, and indeed whole texts. Studies indicate that this approach is likely to work, although more work is needed to validate and optimize it (e.g., compare Mandera, Keuleers, & Brysbaert, 2015, to Hollis & Westbury, in press). If the approach indeed turns out to work, it will be possible to obtain values for all existing words on the basis of a small-scale rating

study. This will be particularly valuable for languages that do not yet have large-scale databases with human ratings.

Indeed, a first important limitation of the current contribution of corpus linguistics is that the measures we discussed are only available for a minority of the 7,000 languages, which does injustice to the language diversity and biases research. Another limitation is that the information is limited to language registers than can easily be analyzed (in particular, written texts). There is an increasing realization that language is inherently multimodal, whereas the corpora are not (yet). This creates a validity problem in relation to the real input for the language user. A solution here might be the creation of multimodal corpora such as the Language archive at the Nijmegen MPI (https://tla.mpi.nl/).

Even for languages that have been included in computational linguistics, another big limitation is that not all measures made available are good or even useful. As it happens, a lot of useless information is to be found on the internet. Using computer algorithms to calculate and compare word features guarantees that one will have a list of numbers as outcome, but does not guarantee that the numbers will be valid. Many things can go wrong. For a start, analyzing big datasets is quite error-prone and requires regular calculation checks. Second, not all algorithms have the same quality (as shown by the research on semantic vectors). Third, much depends on the quality of the corpus one is working with (in this respect it may be good to keep in mind the saying ‘garbage in, garbage out’). Finally, there may be theoretical reasons why the currently used algorithms are suboptimal. For instance, one of the limits of semantic vectors as presently calculated is that antonyms tend to be semantically close on the basis of word co-occurrences. This implies that black is assumed to be a ‘synonym’ of white, and ugly a ‘synonym’ of beautiful.

The best way to avoid bad measures derived from corpus analysis is to validate them against human data. Ideally, this is based on numbers of observations that match those derived

from the corpus. In principle, one could check the usefulness of a new word frequency measure by correlating it to the processing times for some 100 words and see whether it correlates more with the processing times than the prevailing measure, but this is a rather risky strategy, as 100 observations is a small number when the frequency list includes some 100 thousand words. It is much better if one has a database of word processing times for some 20 thousand words. Indeed, research on the quality of word frequency measures and ways to improve them only took off after Balota and colleagues (2007) published a megastudy of lexical decision times (is this letter string a word or not?) and naming latencies for 40 thousand English words. Similarly, it is risky to compare the quality of two semantic similarity measures on the basis of an experiment in which only 20 target words were preceded by related and unrelated primes (as we have done above). The ground is much firmer when one can make use of a megastudy, such as the one by Hutchison et al. (2013), which contains data for 1,661 words preceded by four types of primes.

Megastudies are one source of data for validation studies. They consist of word processing times in popular psycholinguistic tasks (lexical decision, naming, semantic classification, eye movement data). Another interesting source of data consists of human ratings. The best way to test how valid affective estimates based on algorithms are is to compare them to human ratings. Here, again, the size of the database is crucial, so that ratings should be collected for thousands of words. Unfortunately, thus far such sizable databases of human ratings are only available for a few languages (English, Dutch, Spanish). A further use of large databases of human ratings is that they can serve as input for other algorithms, such as those estimating the affective tones of texts (e.g., Hills, Proto, & Sgroi, 2015).

A third interesting validation source is Wordnet (https://wordnet.princeton.edu/). This is a handmade dictionary, available for several languages, in which sets of synonyms (synsets) have been grouped, each expressing a distinct concept, and related to other synsets by a small

number of conceptual relations. In the English database, information is available on 117,000 synsets. The database also contains information about the different meanings and senses of words. For instance, it informs us that 'second' can be used as a noun (with 10 different senses), as a verb (2 senses), an adjective (2 senses), and an adverb (1 sense).

A final human information database that is a useful validation criterion consists of word association data. In word-association studies, participants write down one or more words that come to mind upon seeing or hearing a target word. The standard database up to recently was the Florida Free Association Norms collected in the 1970s and 1980s (http://w3.usf.edu/FreeAssociation/), which contains three-quarters of a million responses to 5,019 stimulus words. An ongoing crowdsourcing study is likely to replace the Florida norms, as it already contains over 4 million responses to 12,000 target words (De Deyne, Navarro, & Storms, 2012; see http://www.smallworldofwords.org/).

There is some irony in the fact that the need for psycholinguistic data is so huge now that corpus linguistics and NLP research produce increasingly better measures of word features (and may soon replace the need for large-scale human ratings). This fact illustrates the everlasting interaction between offline corpus analysis and online human performance research, which is of benefit to both sides.

Key terms

Corpus (corpora): Collection of language produced by humans (speech, written materials, gestures) used to calculate word characteristics, such as word frequency, similarity to other words, and dominant part-of-speech; two important characteristics are the size of the corpus and the representativeness for naturally occurring language.

Corpus linguistics: The study of language through the empirical analysis of large databases of naturally occurring language.

Language register: Variety of language used in a particular setting (e.g., scientific books vs. blogs); is important for psycholinguistics because it has been shown that word characteristics are better at predicting results from experiments if they are based on language participants are likely to have experienced in their life.

Megastudy: Large-scale word processing study in which responses to thousands of words are collected or in which responses from a very large sample of participants are collected; used to examine the variables affecting word processing efficiency and to validate word characteristics calculated in computational linguistics.

Natural language processing (NLP): Discipline that is focused on language processing in computers to increase the interactions with humans, largely based on the analysis of corpora.

Parsing: Syntactic analysis of sentences.

Semantic vector: String of 200-300 numbers describing the meaning of words based on word co-occurrences.

Tagging: Determining the part-of-speech words have in sentences.

Word frequency norms: Estimates of how often words are encountered based on counting their occurrences in representative corpora.

Wordnet: A large lexical database in several languages, in which words have been grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17(9),* 814-823.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445-459.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1)*. Retrieved from http://clic.cimec.unitn.it/marco/publications/acl2014/baroni-etal-countpredict-acl2014.pdf.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Brysbaert, M. & Ellis, A. W. (In press). Aphasia and age-of-acquisition: Are early-learned words more resilient? *Aphasiology*.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology, 2:* 27.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding Part-of-Speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*, 991-997.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016a). 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*, 441-458.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016b). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*:1116. doi: 10.3389/fpsyg.2016.01116.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods, 44*, 890-907.

De Deyne, S., Navarro, D., & Storms, G. (2012). Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations, *Behavior Research Methods, 45*, 480-498.

de Mornay Davies, P. (1998). Automatic semantic priming: The contribution of lexical-and semantic-level processes. *European Journal of Cognitive Psychology, 10*, 389-412.

Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and language, 81*, 162-173.

Fromkin, V. A. (1973) (Ed.) *Speech errors as linguistic evidence*. The Hague: Mouton.

Gimenes, M., & New, B. (In press). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*.

Gries, S. T. (2010). Corpus linguistics and theoretical linguistics A love-hate relationship? Not necessarily... *International Journal of Corpus Linguistics, 15)*, 327-343.

Harris, Z. (1954). Distributional structure. *Word, 10*, 146-162.

Hayes, D.P. (1988). Speaking and writing: Distinct patterns of word choice. *Journal of Memory and Language, 27*, 572-585.

Herdağdelen, A., & Marelli, M. (in press). Social media and language processing: how Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*.

Hills, T. T, Proto, E. & Sgroi, D. (2015), Historical analysis of national subjective wellbeing using millions of digitized books. *IZA Discussion Paper No. 9195*. Retrieved from http://ftp.iza.org/dp9195.pdf.

Hollis, G., Westbury, C., & Lefsrud, L. (In press). Extrapolating human judgments from Skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods, 45*, 1099–1114.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1–37.

Keuleers, E. (2015). *Package 'vwr'*. Retrieved from https://cran.r-project.org/web/packages/vwr/vwr.pdf.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology, 68*, 1457-1468.

Keuleers, E., Diependaele, K. & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology 1:174*. doi: 10.3389/fpsyg.2010.00174.

Francis, W. N., & Kucera, H. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A.B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General, 143*, 1065-1081.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44*, 978-990.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*, 211-240.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185). Retrieved from http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics, 3*. Retrieved from http://u.cs.biu.ac.il/~nlp/wp-content/uploads/Improving-Distributional-Similarity-TACL-2015.pdf

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*, 203-208.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology, 68*, 1623-1642.

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781* [cs]. Retrieved from http://arxiv.org/abs/1301.3781.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General, 118*, 43-71.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language, 57*, 1-23.

Schutz, A. (1962). Common-sense and scientific interpretation of human action. In *Collected Papers I* (pp. 3-47). Springer Netherlands.

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*, 1176-1190.

Further reading and resources

The best textbook on corpus linguistics is Jurafsky D., & Martin, J. H.. (2008). *Speech and language processing* (2nd ed.). Pearson Prentice Hall.. The third edition is foreseen for 2017 (preliminary versions of the chapters can be found on http://web.stanford.edu/~jurafsky/slp3/).

The Language Goldmine website (http://languagegoldmine.com/) includes over 230 links to interesting resources for language research in various languages. Includes most of the links presented here.

The Center for Reading Research website (http://crr.ugent.be/programs-data) includes links to all the variables collected at Ghent University (e.g., word frequency, age of acquisition, concreteness, word prevalence, word valence, arousal), which can be downloaded in various formats. Mostly limited to English and Dutch, however.

Behavior Research Methods (http://www.springer.com/psychology/cognitive+psychology/journal/13428) is the journal in which most word features are published for various languages.

Some of the software packages for corpus research are:

– Natural Language Toolkit (http://www.nltk.org/) – a Python module that provides interfaces to over 50 text corpora and a set of libraries for text processing
– Stanford CoreNLP (http://stanfordnlp.github.io/CoreNLP/) – a set of natural language analysis tools (see also other software released by The Stanford Natural Language Processing Group, http://nlp.stanford.edu/software/index.shtml)
– Gensim (https://radimrehurek.com/gensim/) – a Python module implementing various models used in distributional semantics, including the skip-gram and CBOW models (see also the original word2vec tool released by Google, https://code.google.com/archive/p/word2vec/)

If you want to make use of derived materials, you can use the R package vwr (Keuleers, 2015), download Excel sheets (see above), or make use of websites that allow you to obtain values online. Some of these are:

American English

- http://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus/overview.htm (the SUBTLEX-US database)
- http://elexicon.wustl.edu/ (David Balota's English Lexicon Project)
- http://www.wordfrequency.info/ (Mark Davies's site with word frequencies from various sources)
- http://crr.ugent.be/snaut/ (semantic vectors for English)

British English

- http://crr.ugent.be/archives/1423 (SUBTLEX-UK)
- http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm (slightly dated site with all types of word information)
- http://celex.mpi.nl/ (database with a lot of morphological information)
- http://www.pc.rhul.ac.uk/staff/c.davis/Utilities/ (N-Watch, a program by Colin Davis to obtain various features of English)
- http://crr.ugent.be/programs-data/lexicon-projects (British Lexicon Project, with lexical decisions to 28,000 words)
- http://zipf.ugent.be/snaut/ (semantic vectors for English)

Dutch

- http://crr.ugent.be/isubtlex/ (the SUBTLEX-NL database)
- http://celex.mpi.nl/ (database with a lot of morphological information)
- http://crr.ugent.be/snaut/ (semantic vectors for Dutch)
- http://crr.ugent.be/programs-data/lexicon-projects (Dutch Lexicon Project 1 and 2, with lexical decisions to 30,000 words)

French

- http://www.lexique.org/ (Boris New's site with next to all information about French words)
- https://sites.google.com/site/frenchlexicon/ (the French Lexicon Project with lexical decision times to over 30,000 words)

German

- http://www.dlexdb.de/query/kern/typposlem/ (site with word frequencies in German)
- http://celex.mpi.nl/ (database with a lot of morphological information)

Chinese

- http://crr.ugent.be/programs-data/subtitle-frequencies/subtlex-ch (SUBTLEX-CH word frequencies and PoS information for Chinese words)

Spanish

- http://www.bcbl.eu/databases/espal/ (various word characteristics)
- http://crr.ugent.be/archives/679 (the SUBTLEX-ESP word frequencies)
- http://www.pc.rhul.ac.uk/staff/c.davis/Utilities/ (the N-Watch program for Spanish)