Behav Res
DOI 10.3758/s13428-014-0489-4

Subtlex-pl: subtitle-based word frequency estimates for Polish

Paweł Mandera · Emmanuel Keuleers · Zofia Wodniecka · Marc Brysbaert

Abstract We present SUBTLEX-PL, Polish word frequencies based on movie subtitles. In two lexical decision experiments, we compare the new measures with frequency estimates derived from another Polish text corpus that includes predominantly written materials. We show that the frequencies derived from the two corpora perform best in predicting human performance in a lexical decision task if used in a complementary way. Our results suggest that the two corpora may have unequal potential for explaining human performance for words in different frequency ranges and that corpora based on written materials severely overestimate frequencies for formal words. We discuss some of the implications of these findings for future studies comparing different frequency estimates. In addition to frequencies for word forms, SUBTLEX-PL includes measures of contextual diversity, part-of-speech-specific word frequencies, frequencies of associated lemmas, and word bigrams, providing researchers with necessary tools for conducting psycholinguistic research in Polish. The database is freely available for research purposes and may be downloaded from the authors' university Web site at http://crr.ugent.be/subtlex-pl.

Keywords Word frequencies · Polish language · Lexical decision · Visual word recognition

Electronic supplementary material The online version of this article (doi:10.3758/s13428-014-0489-4) contains supplementary material, which is available to authorized users.

P. Mandera (✉) · E. Keuleers · M. Brysbaert
Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Gent, Belgium
e-mail: pawel.mandera@ugent.be

Z. Wodniecka
Institute of Psychology, Jagiellonian University, Kraków, Poland

Word frequency estimates derived from film and television subtitles have proved to be particularly good at predicting human performance in behavioral tasks. Since lexical decision latencies are particularly sensitive to word frequency (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), correlating human performance in this task with various word frequency estimates became a standard method of validating their usefulness. Word frequencies derived from subtitle corpora were shown to outperform estimates based on written texts for French (New, Brysbaert, Veronis, & Pallier, 2007), English (Brysbaert & New, 2009), Dutch (Keuleers, Brysbaert, & New, 2010), Chinese (Cai & Brysbaert, 2010), Spanish (Cuetos Vega, González Nosti, Barbón Gutiérrez, & Brysbaert, 2011), German (Brysbaert et al., 2011), and Greek (Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010).

Following these developments, we present SUBTLEX-PL, a new set of psycholinguistic resources for Polish, which includes frequency estimates for word forms, associated parts of speech, and lemmas. To our knowledge, this it is the first subtitle word frequency validation study for a Slavic language. In terms of number of speakers, Polish is the largest language in the West Slavic group and the second largest of all Slavic languages after Russian (Lewis, Simons, & Fennig, 2013). It is a highly inflected language and, as compared with most Germanic languages, has a much richer inflection of nouns, adjectives, verbs, pronouns, and numerals. Polish is written in the Latin alphabet, with several additional letters formed with diacritics. In contrast to English, Polish has a transparent orthography: In most cases, letters or their combinations correspond to phonemes of spoken Polish in a consistent way.

Even though the collection of text corpora of considerable size is easier than ever before, the standard way of validating the quality of the word frequencies based on these corpora has typically involved collection of data for thousands of words in

Published online: 19 June 2014

strictly controlled laboratory settings (Balota et al., 2007; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2011). In order to compare frequency estimates derived from two corpora, it may be more efficient to use words for which the two corpora give diverging estimates, rather than a random set of words. This idea is based on the observation that the words for which the frequency estimates between two corpora differ most are also the sources of potential difference in performance of these frequency norms when predicting behavioral data. This approach can increase the statistical power of the experiment; if only randomly sampled words are included in the study, due to very high correlation between different frequency estimates, it is more difficult to detect differences in performance of these estimates without including a very large number of words in the experiment. Dimitropoulou et al. (2010) approached this problem by using a factorial design in which the critical conditions included words with a high frequency in one corpus and a low frequency in the other. In the present study, we will use an approach based on continuous sampling over the full range of word frequencies.

Although using words for which the two corpora give the most diverging estimates may help to detect differences between their performance in predicting behavioral data, there is a possibility that this approach may bias the experiment in favor of one of the frequency estimates. For instance, words in the formal register tend to have a much higher frequency in written corpora than in spoken corpora. Stimulus selection based solely on a criterion of maximum divergence would lead to a large selection of words from the formal register, while the formal register may represent just a small part of the corpus. To account for this possibility, in Experiment 1, we included an additional set of words that were randomly sampled from all word types observed in the compared corpora. In Experiment 2, we included only randomly sampled words.

Current availability of frequency norms for Polish

For a long time, the only available word frequency norms for Polish were based on a corpus compiled between 1963 and 1967 (containing about 500,000 words) and published by Kurcz, Lewicki, Sambor, Szafran, and Woroniczak (1990). More recently, several other Polish text corpora have been compiled, and resources such as concordances and collocations have been made available to researchers. This is the case for the IPI PAN Corpus of about 250 million words (Przepiórkowski & Instytut Podstaw Informatyki, 2004), the Korpus Języka Polskiego Wydawnictwa Naukowego PWN (n.d.), containing about 100 million words, and the PELCRA Corpus of Polish (~100 million words; http://korpus.ia.uni.lodz.pl/). To our knowledge, none of them provides an easily accessible list of word frequencies.

The largest of the Polish corpora contains over 1.5 billion words (National Corpus of Polish [NCP]; Przepiórkowski, 2012). It is based mainly on press and magazines (~830 million tokens), material downloaded from the Internet (~600 million tokens), and books (~100 million tokens). It also contains a small sample of spoken, conversational Polish (~2 million tokens). In addition to the full corpus, a significant effort has been invested in creating a subcorpus that is representative of the language exposure of a typical native speaker of Polish. This balanced subcorpus (BS–NCP) contains about 250 million words. Spoken materials (conversational and recorded from media) constitute about 10 % of the subcorpus. The remaining 90 % is based on written texts (mainly from newspapers and books).

Since the word frequencies derived from the NCP balanced subcorpus seem to be the most appropriate existing word frequencies for psycholinguistic research in Polish, we decided to compare them with the new SUBTLEX-PL frequencies.

SUBTLEX-PL

Corpus compilation, cleaning, and processing

We processed about 105,000 documents containing film and television subtitles flagged as Polish by the contributors of http://opensubtitles.org. All subtitle-specific text formatting was removed before further processing.

To detect documents containing large portions of text in languages other than Polish, we first calculated preliminary word frequencies on the basis of all documents and then removed from the corpus all files in which the 30 most frequent types did not cover at least 10 % of a total count of tokens in the file. Using this method, 5,365 files were removed from the corpus.

Because many documents are available in multiple versions, it was necessary to remove duplicates from the corpus. To do so, we first performed a topic analysis using Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), assigning each file to one of 600 clusters. If any pair of files within a cluster had an overlap of at least 10 % unique word-trigrams, the file with the highest number of hapax legomena (words occurring only once) was removed from the corpus, since more words occurring once would indicate more misspellings.

After removing duplicates, 27,767 documents remained, containing about 146 million tokens (individual strings, including punctuation marks, numbers, etc.), out of which 101 million tokens (449,300 types) were accepted as correctly spelled Polish words by the Aspell spell-checker (http://aspell.net/; Polish dictionary available at ftp://ftp.gnu.org/gnu/aspell/dict/pl/) and consisted only of legal Polish, alphabetical characters. All words were converted to lowercase before spell-checking. Because Aspell rejects

proper names spelled with lowercase, this number does not include proper names.

Frequency measures

Word frequency

In addition to raw frequency counts, it is useful for researchers to have measures of word frequency that are independent of corpus size. First, we report word frequencies transformed to the Zipf scale[1] (van Heuven, Mandera, Keuleers & Brysbaert 2014). The Zipf scale was proposed as a more convenient scale on which word frequencies may be measured. In order to reflect the nature of the frequency effect, it is a logarithmic scale (like the decibel scale of sound intensity), but, in contrast to the logarithm of frequency per million words, it does not result in negative values for corpora of up to 1 billion words. In order to make interpretation of the frequency values easier, the middle of the scale separates low-frequency from high-frequency words, and, for a majority of words, the measure takes a value between 1 to 7, which resembles a Likert scale. Another compelling property of the Zipf scale is that it allows assigning a value to words that were not observed in a corpus by incorporating Laplace smoothing, as recommended by Brysbaert and Diependaele (2013); without the transformation, such words pose a problem, since the logarithm of 0 is undefined, which makes it impossible to estimate $\log_{10}$ of word frequency per million for these words. In addition to the raw frequency and the Zipf scale frequencies, we also provide the more traditional logarithm of frequency per million words.

Contextual diversity

Adelman, Brown, and Quesada (2006) proposed that the number of contexts in which a word appears may be more important than word frequency itself and that the number of documents in which a word occurs may be a good proxy measure for the number of contexts (contextual diversity [CD]). According to this view, even words with equal frequency would be processed faster if they occur in more contexts. Brysbaert and New (2009) observed that CD accounts for 1 %–3 % more variance than does word frequency.

Part-of-speech-specific frequencies

For languages with a rich inflectional system, such as Polish, it is crucially important to provide researchers with information above the level of individual word forms. For each word in SUBTLEX-PL, we also provide the lemma and the dominant part of speech and their frequencies.

Providing the lemma associated with each given word form allows us to group inflected forms of the same word. This may be useful when investigating the specific contributions of surface and lemma frequencies in word processing (Schreuder & Baayen, 1997) or in order to avoid including inflections of the same word when creating a stimulus set for an experiment.

Information about the dominant part of speech allows researchers to choose words of a particular grammatical class (e.g., when a researcher wants to include only nouns in a stimulus list).

To obtain part-of-speech and lemma information for words, we used TaKIPI, a morphosyntactic tagger for Polish (Piasecki, 2007) supplied with the morphological analyzer Morfeusz (Woliński, 2006). The resulting tag set was too detailed for our purposes, so we translated the original tags to a simpler form that includes only information about parts of speech and discards other details.[2] The tagging process assigned each of the word forms consisting of legal Polish alphabetical characters and accepted by the spell-checker to 1 of 78,361 lemmas.

Bigram frequencies

Although in this article we focus on unigram frequencies, we also provide frequency estimates for word bigrams, which are of increasing interest to researchers (Arnon & Snider, 2010; Siyanova-Chanturia, Conklin, & van Heuven, 2011).

Experiment 1

Method

Stimuli

We selected stimuli from the list of words common to both BS–NCP and SUBTLEX-PL.[3] All stimuli considered for selection contained only alphabetical characters and occurred without an initial capital in most cases. We used the list of 1-grams (available at http://zil.ipipan.waw.pl/NKJPNGrams) to generate the BS–NCP frequency list used in the present study. We processed the raw list by summing frequencies of all forms that were identical after removing punctuation marks attached to some of the forms in the original list.

[1] $z_i = \log_{10}(\frac{c_i+1}{\sum_{k=1}^{n} c_k+n}) + 9$ (van Heuven, Mandera, Keuleers & Brysbaert 2014) Where z_i is a Zipf value for word i, c_i is its raw frequency, and n is the size of the vocabulary.

[2] For mapping between original and simplified tags, see supplementary materials.

[3] A nonfinal version of SUBTLEX-PL, based on nearly 50 million tokens, was used when choosing stimuli for the experiment.

To make the experiment maximally informative, we chose stimuli for which BS–NCP and SUBTLEX-PL gave highly divergent frequency estimates. We performed a linear regression on the SUBTLEX-PL frequencies, using the BS–NCP frequencies as a predictor. All frequencies were transformed to the Zipf scale. We then ordered the words according to their residual error and chose 155 words from both extremes of the resulting list, ensuring that different forms of the same lemma were not selected more than once. Words at one extreme (with a large positive residual error value) were much more frequent in SUBTLEX-PL than would be expected on the basis of BS–NCP, while words at the other extreme (with a large negative residual error value) occurred much less often in SUBTLEX-PL than would be expected on the basis of BS–NCP. In addition, we randomly sampled 155 words from the remaining words, with the probability of each word being selected equal to its probability in the subtitle corpus.

Figure 1 illustrates the frequency distribution of stimuli according to this procedure. As the top panel of Fig. 1 shows, it is important to note that the regression line on which the residual error values are based is pulled downward by a large number of words with a low frequency in SUBTLEX-PL. While this seems to indicate that SUBTLEX-PL contains a higher proportion of low-frequency forms, it is an artifact of selecting words from corpora of unequal size.[4]

Words that had a much higher frequency in one corpus than in the other may be categorized into several groups. For example, words related to the Polish administrative and legislative system (e.g., "województwo," *district*; "urzędowym," *administrative*), as well as those occurring mostly in fairly sophisticated contexts (e.g., "pejzażu," *landscape*) are much more frequent in the BS–NCP corpus. On the other hand, words with much higher frequency in SUBTLEX-PL included those used mostly in dialogues (e.g., "skarbie," *honey*), swear words ("pierdol," *fuck*), those related to (American) film themes (e.g., "kowboju,"*cowboy*), and function words (e.g., "ale," *but*; "się," *self*).

For each word that was included in the experimental set, a corresponding nonword was generated using Wuggy, a multilingual pseudoword generator (Keuleers & Brysbaert, 2010).

For the full set of words included in the experiment, the standard deviation (*SD*) in word frequency (Zipf scale) was 1.14 (mean = 4.09) for BS–NCP and 1.76 (mean = 3.63) for SUBTLEX-PL. The two variances were significantly different, $F(464, 464) = 0.42$, $p < .001$, and Welsch's *t*-test has shown significant differences in the mean frequency derived from the two corpora, $t(794) = 4.7$, $p < .001$, for this set of stimuli.

For the 155 word stimuli that were randomly sampled from the words common to both word frequency lists, *SD* was 1.08 (mean = 4.44) for BS–NCP and 1.19 (mean = 4.11) for SUBTLEX-PL. The difference between variances was not statistically significant, $F(154, 154) = 0.82$, $p = .23$, but the mean frequencies were significantly different according to Welsch's *t*-test, $t(308) = 2.6$, $p = .01$.

Participants

Twenty-six students from the Jagiellonian University in Kraków participated in the experiment (20 female, 6 male; mean age=23.76, *SD* = 2.06) either on a voluntary basis or in exchange for course credit.

Design

Words and nonwords were randomly assigned to 10 blocks. Nine blocks contained 50 words and 50 nonwords in a random order; 1 block contained the remaining 15 words and nonwords in a random order. Ten different permutations of block orders were generated, and each participant was randomly assigned to one of the permutations. Due to a coding error, 10 words were not presented to the first 10 participants. Further analysis is therefore based on 455 words, instead of 465 words.

Within each block, stimuli were presented in a random order in white characters on a black background. Presentation of each stimulus was preceded by a blank screen. After 500 ms, a vertical line was displayed above and below the center of the screen. Finally, after another 500 ms, the stimulus was presented between the vertical lines.

A standard QWERTY PC keyboard was used to collect responses. Participants were instructed to press "/" (the rightmost key on the second row) if they saw a word and "Z" (the leftmost key on the second row) if they saw a nonword. The time-out for giving the response was 2,000 ms. After six training trials, the experimental blocks were presented. The experiment took about 30 min.

Results

Of the trials on which reaction times (RTs) were outside of a range of whiskers of a boxplot adjusted for skewed distributions (calculated separately for words and nonwords for each participant in each block; Hubert & Vandervieren, 2008), 5.2 % were removed from the data set.

Accuracy and RTs were the two dependent variables in all analyses. Three stimuli with less than one-third correct answers were excluded from the data set. The analyses are reported first for the full set of words included in the

[4] As an example, consider a list of 200,000 words and a list of 400,000 words. A typical characteristic of word frequency distributions is that about half of the words in each list will have a frequency of one. In that case, the base probability that any word found in both lists would have a frequency of 1 in the first list would be 1/100,000, while it would be 1/200,000 for the second list.

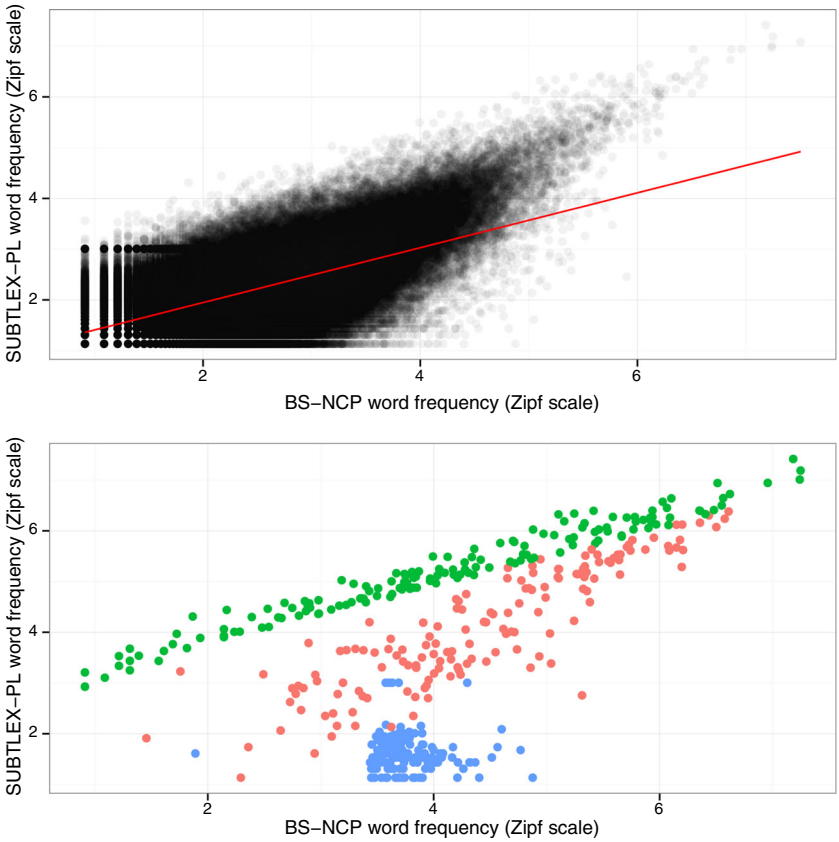

Fig. 1 Frequencies of words in the BS–NCP and SUBTLEX-PL corpora for all words (upper panel; the red line shows a regression line predicting SUBTLEX-PL frequencies based on BC–NCP frequencies) and words included in Experiment 1 (bottom panel) showing randomly sampled words (red) and words with higher frequency (green) and lower frequency (blue) in SUBTLEX-PL than in BS–NCP

experiment and then separately only for the 155 word stimuli that were randomly sampled from the words common to both word frequency lists.

For the full set of word stimuli, the mean RT was 592.00 (*SD* = 67.34), and the mean accuracy was .94 (*SD* = .08). Words occurring less often in SUBTLEX-PL than in BS–NCP had a mean RT of 652.19 (*SD* = 52.23) and a mean accuracy of .96 (*SD* = .06), while words occurring more often in SUBTLEX-PL than in BS–NCP had a mean RT of 551.02 (*SD* = 48.74) and a mean accuracy of .91 (*SD* = .11). The randomly selected words had a mean RT of 574.00 (*SD* = 54.00) and a mean accuracy of .96 (*SD* = .07).

For nonwords, the mean RT was 666.88 (*SD* = 70.23), and the mean accuracy was .94 (*SD* = .09).

To estimate the reliability of the RT and accuracy measures, we computed split-half correlations for 100 random splits of the data across participants. The resulting correlations were corrected with the Spearman–Brown prediction formula (Brown, 1910; Spearman, 1910), giving an average corrected reliability of .81 (*SD* = .013) for RTs and .72 (*SD* = .021) for accuracy.

Adjusted R^2 was used as a measure of explained variance in all analyses. The percentage of variance in RT and accuracy accounted for by linear regression models using different frequency measures is summarized in Table 1. All frequency measures were transformed to the Zipf scale (van Heuven et al., 2014). Because it was shown that the frequency effect is not completely linear (Balota et al., 2004), we added a term with squared word frequency (Zipf scale) to the linear regression. To control for word length, we also included number of letters in a word in the regression model.

The relationship between word frequencies and RTs is shown in Fig. 2. As is shown in Table 1, when all words were included in the analysis, the BS–NCP word frequencies explained 39.09 % of variance in RTs and 8.90 % of variance in accuracy. For this set of words, SUBTLEX-PL frequencies explained 58.64 % of variance in RTs and 19.07 % in accuracy, which is 19.55 % more for RTs and 10.17 % more for accuracy in comparison with BS–NCP frequencies. To test for statistical difference between models, we applied the Vuong test for nonnested models (Vuong, 1989). The differences in performance of the two models were statistically significant for both RTs ($z = -6.11$. $p < .001$) and accuracy ($z = -2.5$, $p = .012$).

When only words that were randomly sampled from the corpus were included in the analysis, the frequencies

Table 1 Percentages of variance accounted for by the various frequency measures in Experiment 1

Model	RT (%; all words)	Accuracy (%; all words)	RT (%; sampled words)	Accuracy (%; sampled words)
$Length + WF_{BS-NCP} + WF_{BS-NCP}^2$	39.09	8.90	45.53	20.58
$Length + WF_{SUB-PL} + WF_{SUB-PL}^2$	58.64	19.07	53.88	18.43
$Length + CD_{SUB-PL} + CD_{SUB-PL}^2$	59.72	20.81	54.35	19.26
$Length + WF_{SUB-PL} + WF_{SUB-PL}^2 + DLF$	58.80	20.16	53.59	18.52
$Length + CD_{SUB-PL} + CD_{SUB-PL}^2 + DLF$	59.77	21.64	54.10	19.20
$Length + WF_{SUM} + WF_{SUM}^2$	50.99	19.14	51.01	22.01
$Length + WF_{AVG} + WF_{AVG}^2$	58.36	21.38	55.46	21.77

Note. Columns 2 and 3 show the results for all words in the experiment; columns 4 and 5 show the results for randomly sampled words. WF = word frequency (Zipf scale), DLF = log_{10} of dominant lemma frequency, BS–NCP =Balanced Subcorpus–National Corpus of Polish, SUB-PL = Polish Subtitle Corpus, WF_{SUM} = normalized (Zipf scale) sum of word frequencies in SUBTLEX-PL and BS–NCP, WF_{AVG} = averaged Zipf scale frequency in the two corpora

derived from the BS–NCP corpus explained 45.53 % of the variance in RTs and 20.58 % in accuracy. In this case, the difference between the BS–NCP and SUBTLEX-PL corpora was smaller, and word frequencies derived from the SUBTLEX-PL corpus explained 8.35 % more variance for RTs but 2.15 % less variance for accuracy. The difference was not significant for RTs ($z = -1.84$, $p = .065$) or accuracy ($z = 0.45$, $p = .65$).

For the full set of words, CD measures calculated on the basis of SUBTLEX-PL accounted for the largest part of the variance for both RTs and accuracy, explaining 59.72 % and 20.81 % of variance, respectively. This improvement of model

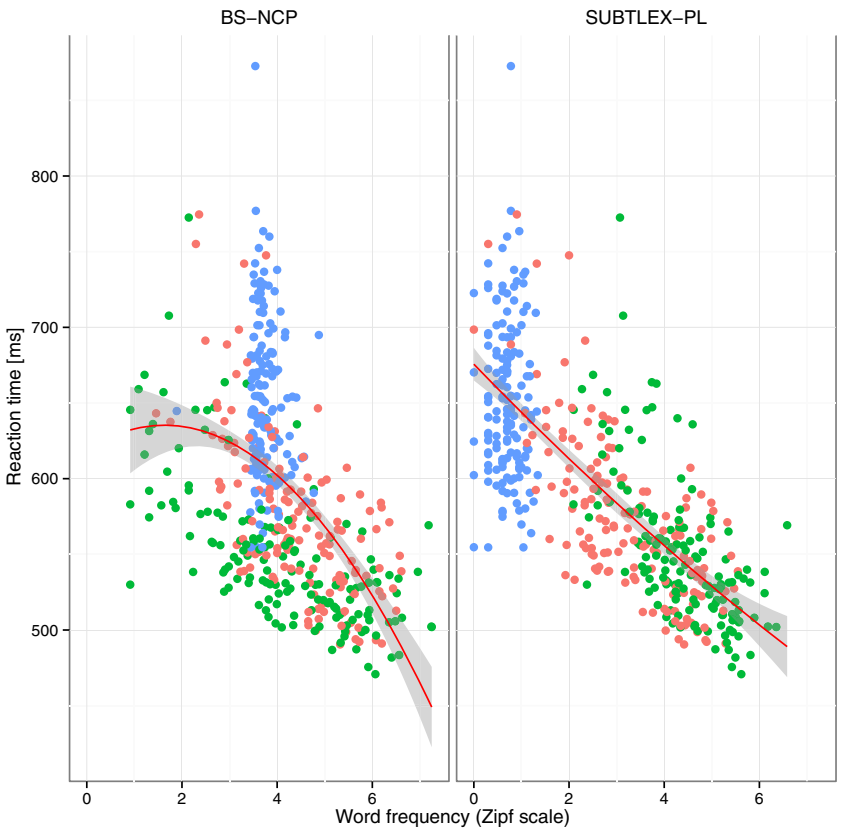

Fig. 2 Reaction times in Experiment 1 for words and their frequencies in the BS–NCP (left) and SUBTLEX-PL (right) corpora. Reaction times for words that had much higher frequencies in BS–NCP, as compared with SUBTLEX-PL (blue), are shifted upward from the regression line, while words that have higher frequencies in SUBTLEX-PL than in BS–NCP (green) tend to be responded to faster than would be predicted on the basis of BS–NCP frequencies. Reaction times predicted on the basis of SUBTLEX-PL line up much closer to the regression line. For words that were randomly sampled from the full set of words (red), this difference is less apparent, but it is still reflected in R^2. Red lines represent predictions of a linear model with word frequency and its square term as predictors (with standard error in the shaded area)

predictions, relative to the one based on word frequencies, was statistically significant for both RTs ($z = 2.41$, $p = .016$) and accuracy ($z = 2.57$, $p = .010$).When only randomly selected words were included in the analysis, CD explained 54.35 % of variance for RTs and 19.26 % for accuracy. This was not significantly better than the model based on subtitle word frequencies for RTs ($z = 0.86$, $p = .39$) or for accuracy ($z = 1.15$, $p = .25$).

To examine the importance of lemma frequency, we conducted further analyses including dominant lemma frequency as an additional predictor. This predictor turned out to add very little to the total amount of explained variance. The Vuong test has not indicated in any case that the model including this predictor should be preferred over a simpler model.

In addition to analyses based on frequencies derived from SUBTLEX-PL and BS–NCP, we also calculated compound measures of word frequency, taking into account frequencies in the two corpora simultaneously: their summed frequency (transformed to the Zipf scale after summation) and their averaged normalized (Zipf scale) frequency. In the case of the full set of word stimuli, in comparison with BS–NCP frequencies, the summed frequency measure explained 11.89 % more variance in RTs ($z = 6.38$, $p < .001$) and 10.24 % more variance in accuracy ($z = 2.97$, $p = .003$). In comparison with subtitle frequencies, it explained 7.66 % less variance in RTs ($z = -2.93$, $p = .003$) and a similar amount of variance in accuracy ($z = 0.016$, $p = .99$). The averaged frequency explained 7.3 % more variance in RTs than did the summed frequency ($z = 4.40$, $p < .001$) and a comparable amount of variance to subtitle frequencies ($z = -0.16$, $p = .87$). For accuracy, its predictions were not significantly better than summed frequencies ($z = 0.84$, $p = .40$) or subtitle frequencies ($z = 1.03$, $p = .30$) and outperformed only BS–NCP-based frequencies (by 12.50 % of explained variance; $z = 4.157$, $p < .001$).

For a randomly sampled set of words, the compound measures performed particularly well: The model using estimates based on averaged normalized frequency in the two corpora accounted for 1.1 % more variance in RTs than did the next best model (based on SUBTLEX-PL contextual diversity), but the difference between the two models was not statistically significant ($z = 0.38$, $p = .70$). In comparison with the model based on BS–NCP word frequencies, both summed frequency ($z = 2.86$, $p = .004$) and averaged frequency ($z = 3.65$, $p < .001$) performed significantly better in predicting RTs. As compared with the model based on SUBTLEX-PL frequencies, the difference was not statistically significant for either of the compound measures (for summed word frequency, $z = -0.073$, $p = .46$; for averaged word frequency $z = 0.57$, $p = .57$). The two compound measures were also best at predicting accuracy, but none of the differences in accuracy reached the level of statistical significance ($z < 1.96$).

Discussion

In Experiment 1, we found a general advantage of SUBTLEX-PL frequencies. The difference was larger when stimuli with extremely divergent frequency estimates were included in the analyzed data set. At first sight, these results suggest that the SUBTLEX-PL word frequencies are more balanced than the BS–NCP word frequencies: RTs for the three different groups of stimuli are in line with the predictions from SUBTLEX-PL. On the other hand, the BS–NCP frequencies seem to severely underestimate RTs for words that have a much lower occurrence in SUBTLEX-PL (shown in blue in Fig. 2). This could indicate that the BS–NCP corpus has inflated frequency estimates for these words, of which most could be characterized as belonging to a very formal register.

However, we should note that the frequency range of the sample of words for which BS–NCP makes the worst predictions is very restricted, making a general conclusion about the global suitability of the BS–NCP frequencies premature. Researchers will not often encounter a situation where an experiment requires exactly this register of words. Moreover, when only randomly sampled words were included in the data set, the difference between performance of the two frequency estimates was smaller, and the advantage of SUBTLEX-PL was no longer statistically significant.

In additional analyses, we have shown that compound frequency estimates, taking into account both corpora simultaneously, can be particularly good predictors of performance in a lexical decision task. This can be due to the fact that considering the two corpora simultaneously involves a significant increase in the overall size of a sample of a language on which frequency estimates are based. In addition to that, compounding word frequency estimates may help reduce bias for certain registers that may be present in the individual corpora.

In Experiment 2, we propose a comparison of the two word frequency measures in which the entire frequency distribution is examined and undue bias from a particular register is avoided.

Experiment 2

Method

Participants

For the second experiment, 43 female participants and 15 male participants took part in an online experiment. Mean age of the participants was 27.07 ($SD = 4.08$; 1 of the participants did not give information about age).

Stimuli

Three hundred word stimuli were selected using a two-step sampling procedure. First, simple Good-Turing Smoothing (e.g., Gale & Sampson, 1995) was applied to the word frequencies from BS–NCP and SUBTLEX-PL (Brysbaert & Diependaele, 2013). Words that were present in both word frequency lists and had a length of at least three letters were considered for further selection if they were included in the PWN dictionary (http://sjp.pwn.pl). The probability of a word being selected for the experiment was proportional to its simple Good-Turing Smoothed probability, averaged over BS–NCP and SUBTLEX-PL. Once a word had been selected, other words forms of the same lemma were ignored, avoiding including different inflections of the same word in the stimulus list. Three hundred nonwords were generated using Wuggy (Keuleers & Brysbaert, 2010) on the basis of an independent sample of words from the SUBTLEX-PL and BS–NCP corpora.

Figure 3 shows the relationship between the BS–NCP and SUBTLEX-PL word frequencies for the stimuli in Experiment 2. Standard deviation in word frequency (Zipf scale) was 1.46 (mean=3.81) for BS–NCP and 1.59 (mean=3.72) for SUBTLEX-PL. There were no statistically significant differences between frequencies derived from the two corpora in means (Welsh's *t*-test), $t(594) = -0.74$, $p = .46$, or their variances, $F(299, 299) = 1.2$, $p = .14$.

Design

The experiment was administered in a Web browser, using custom-designed software, taking into account timing (Crump, McDonnell, & Gureckis, 2013). Participants were instructed to respond by pressing "J" if they thought that the presented stimulus was a word and "F" if they thought that it was not a word. After a short training block with 4 words and 4 nonwords, during which feedback was given after each trial, experimental stimuli were presented in five blocks. For each block, 60 words and 60 nonwords were chosen at random. After each block, feedback was given about performance (mean RT for words and overall accuracy in the preceding block). Participants were allowed to take a short break between blocks. Stimuli were presented in black font on a white background until the participant gave a response, after which the screen would be blank for 500 ms before the next stimulus was displayed. During the experiment, a continuous progress bar was presented in the upper part of the screen.

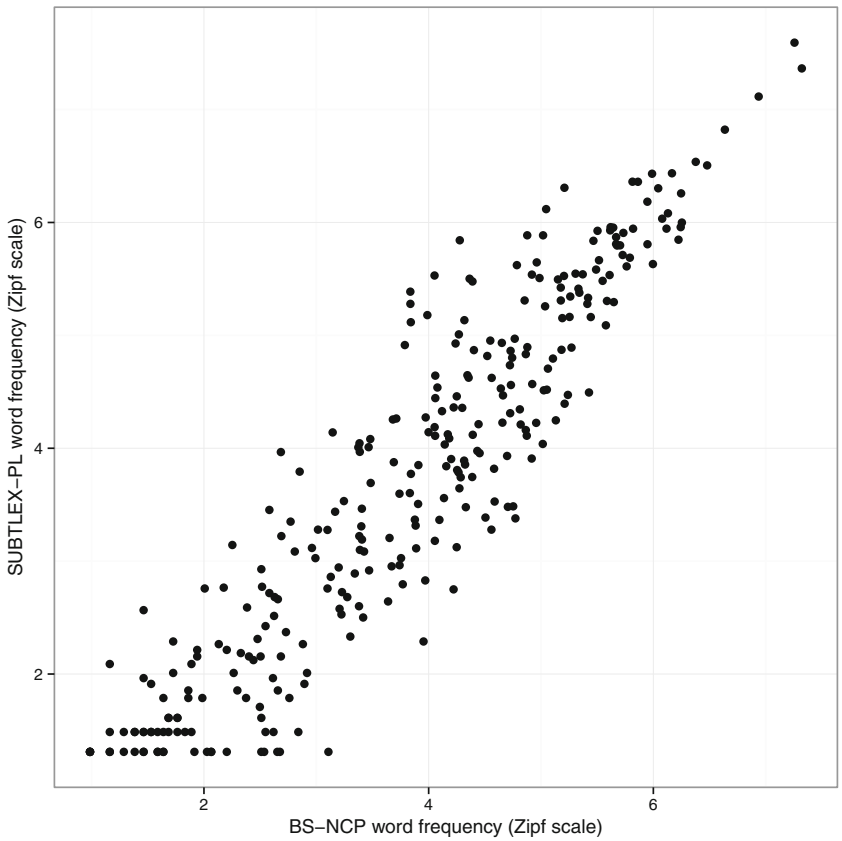

Fig. 3 Frequencies in the BS–NCP and SUBTLEX-PL corpora for words included in Experiment 2

Results

To exclude outliers from the analyzed data set, a two-step procedure was applied. First, we excluded all trials with RTs longer than 3,000 ms. Next, all observations in which RTs were outside a range of whiskers of a boxplot adjusted for skewed distributions (calculated separately for words and nonwords for each participant in each block; Hubert & Vandervieren, 2008) were removed from the data set. In total, 8 % of trials were removed.

The mean accuracy was .96 for words and .97 for nonwords. Mean RT was 893.97 (*SD* = 188.03) for words and 1,043.79 (*SD* = 174.63) for nonwords. On average, the RTs were substantially longer than in the first experiment, most likely because of the lack of a time-out and the fact that most participants in Experiment 1 were used to taking experiments for course credit.

Reliability of the RT and accuracy measures was computed in the same way as for Experiment 1. The mean corrected reliability was .94 (*SD* = .005) for RTs and .88 (*SD* = .013) for accuracy.

In Experiment 2, as compared with SUBTLEX-PL frequencies, the BS–NCP frequencies accounted for 2.4 % more variance in RTs and for 3 % more variance in accuracy (see also Table 2 and Fig. 4); however, the difference in performance of the two models was not statistically significant for RTs ($z = 1.12$, $p = .26$) or for accuracy ($z = 1.00$, $p = .32$). The compound frequency estimates turned out to give the most accurate predictions of RTs. Although, in comparison with the model based on BS–NCP word frequencies, this difference was not statistically significant for summed frequencies ($z = 1.49$, $p = .14$) or for averaged frequencies ($z = 0.83$, $p = .40$), in comparison with the model based on movie subtitles, both compound measures performed significantly better: The summed frequencies explained 3.4 % more variance ($z = 2.02$, $p = .043$) and averaged frequencies 3.2 % more variance ($z = 2.66$, $p = .008$) in RTs. The model, which included dominant lemma frequencies in addition to subtitle frequencies, significantly outperformed the model without this predictor ($z = 2.11$, $p = .035$).

For accuracy, the measures derived from BS–NCP followed these based on SUBTLEX-PL contextual diversity and dominant lemma frequency in explained percentage of the variance. None of the differences in accuracy reached the level of statistical significance ($z < 1.96$).

Table 2 Percentages of variance accounted for by the various frequency measures in Experiment 2

Model	RT (%)	Accuracy (%)
$Length + WF_{BS-NCP} + WF_{BS-NCP}^2$	70.48	19.05
$Length + WF_{SUB-PL} + WF_{SUB-PL}^2$	68.06	16.02
$Length + CD_{SUB-PL} + CD_{SUB-PL}^2$	68.32	17.4
$Length + WF_{SUB-PL} + WF_{SUB-PL}^2 + DLF$	70.71	18.96
$Length + CD_{SUB-PL} + CD_{SUB-PL}^2 + DLF$	70.72	19.55
$Length + WF_{SUM} + WF_{SUM}^2$	71.45	18.37
$Length + WF_{AVG} + WF_{AVG}^2$	71.31	18.51

Note. WF = word frequency, BS–NCP = Balanced Subcorpus–National Corpus of Polish, SUB-PL = Polish Subtitle Corpus

Discussion

In Experiment 2, the compound measures again performed best in predicting behavioral data. Interestingly, for models based on frequency estimates derived from BS–NCP and SUBTLEX-PL, we observed a reversed pattern, relative to Experiment 1: The SUBTLEX-PL frequencies were now worse at predicting RTs, as compared with the compound measures, but this was not the case for BS–NCP frequencies. Even more surprisingly, the randomly sampled words in Experiment 1 showed the reverse pattern. We suspected that this was caused by different means and standard deviations in frequencies between the two experiments. The average frequency was higher in the first experiment (for both corpora) than in the second experiment. Hence, the two corpora may differ in their potential to explain variance in RTs in various frequency ranges. To test this hypothesis, we performed an additional analysis using a linear regression model with number of letters, word frequency in BS–NCP, word frequency in SUBTLEX-PL, and the interaction between the frequencies of both corpora. Table 3 shows the results of this analysis. Because the interaction between the two frequency measures turned out to be highly significant, we decided to conduct an additional analysis. We split the set of words in Experiment 2 at the median point of average word frequency in the two corpora (3.8, Zipf scale). We observed (see Table 4) that the BS–NCP frequencies are better in predicting RTs and accuracy in the lower part of the frequency range, while SUBTLEX-PL frequencies are better in predicting these variables in the higher part of the frequency range. The difference in performance of the models based on frequencies derived from individual corpora was not significant in the upper part of the frequency range ($z = 1.72$, $p = .086$) or in the lower part of the frequency range ($z = 1.34$, $p = .18$), but the model based on averaged frequencies was best in both frequency ranges. It significantly outperformed BS–NCP-based frequencies in the higher range ($z = 2.34$, $p = .019$) and the model based on subtitle frequencies in the lower range ($z = 2.03$, $p = .042$). For accuracy, the Vuong test did not show preference for any of the models ($z < 1.96$).

In order to verify whether a similar interaction between frequency estimates derived from primarily written-text and subtitle-based corpora can be found in other languages, we conducted an additional analysis using RTs collected in the British Lexicon Project (BLP; Keuleers et al., 2011). We used

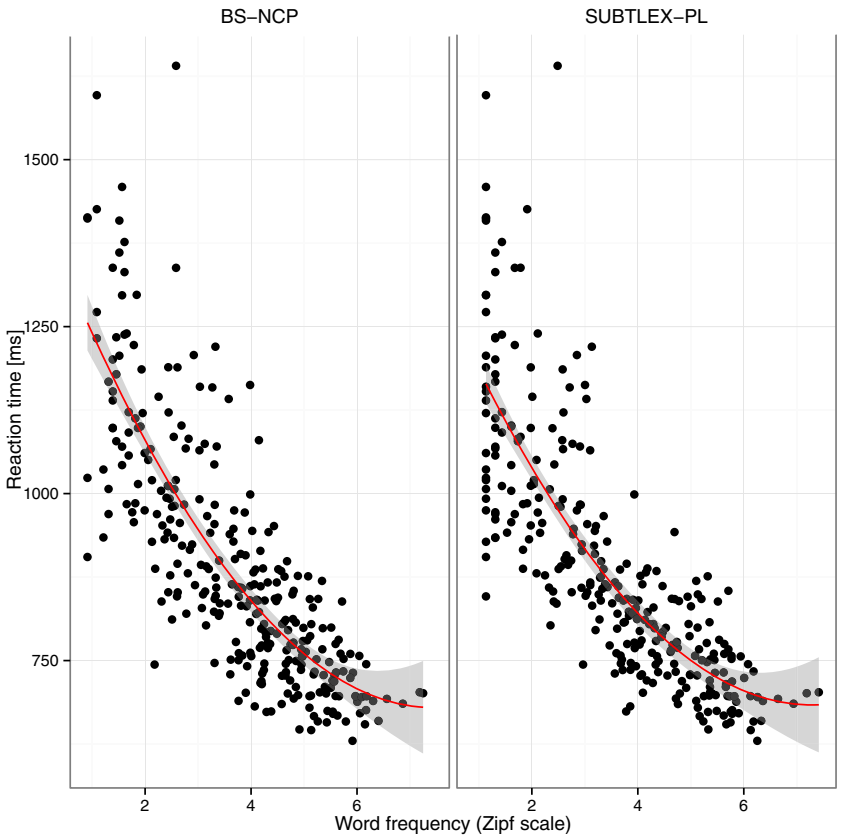

Fig. 4 Reaction times for words and their frequencies in the BS–NCP (left) and SUBTLEX-PL (right) corpora. The red lines represent predictions of a linear model with word frequency and its square term

frequency estimates from the British National Corpus (BNC; Kilgarriff, 2006), which consists mostly of written language and contains about 100 million words, and SUBTLEX-UK (van Heuven et al., 2014). To emulate the setup of the experiment reported in the present article and to better balance the number of words from different frequency ranges, we ran 1,000 simulations in which we randomly chose 300 words from the BLP with weights proportional to the averaged word frequency (Zipf scale) of the BNC and SUBTLEX-UK. For each sample, we fitted a linear model with number of letters, word frequency in the BNC, word frequency in SUBTLEX-UK, and the interaction between the word frequencies of both corpora.

We found that the interaction between the two frequency measures was highly significant ($p < .001$) in all 1,000 simulations. At the same time, we did not find an advantage of BNC word frequencies in the lower part of the frequency spectrum when the stimuli in each of the samples was split at the median point (mean median point=3.21, $SD = 0.061$, Zipf scale). Across all the samples, in the lower part of the

Table 3 Regression model for predicting reaction times using length of a word, frequencies derived from BS–NCP and SUBTLEX-PL, and interaction term between the two corpora

	Estimate	*SE*	*t*-value	*p*
Intercept	772.67	21.07	36.67	<2e-16
Length	15.46	2.58	5.99	6.1e-09
$WF_{BS\text{-}NCP}$	−105.87	13.28	−7.97	3.5e-14
WF_{SUB-PL}	−101.47	14.75	−6.88	3.6e-10
$WF_{SUB-PL} * WF_{BS\text{-}NCP}$	17.05	2.61	6.54	2.7e-10

Adjusted $R^2 = .71$; $F(4, 295) = 186.00$, $p < 2e\text{-}16$

Note. The frequencies were centered before being entered into the linear regression

Table 4 Percentage of variance explained by frequency estimates derived from the two corpora (the data set from Experiment 2 was split at the median)

Frequency	Model	RT (%)	Accuracy (%)
> median	$length + WF_{BS-NCP} + WF_{BS-NCP}^2$	27.49	9.65
> median	$length + WF_{SUB-PL} + WF_{SUB-PL}^2$	33.89	11.72
> median	$length + WF_{SUM} + WF_{SUM}^2$	31.89	11.63
> median	$length + WF_{AVG} + WF_{AVG}^2$	33.79	12.17
≤ median	$length + WF_{BS-NCP} + WF_{BS-NCP}^2$	45.70	14.05
≤ median	$length + WF_{SUB-PL} + WF_{SUB-PL}^2$	38.38	12.92
≤ median	$length + WF_{SUM} + WF_{SUM}^2$	46.20	13.89
≤ median	$length + WF_{AVG} + WF_{AVG}^2$	45.45	14.38

Table 5 Regression model for predicting reaction times in Experiment 2 using word length , word frequency (WF_{SUB-PL}), log_{10} of dominant lemma frequency (DLF), and the interaction between form and lemma frequencies

	Estimate	*SE*	*t*-value	*p*
Intercept	787.87	20.69	38.09	<2e-16
Length	13.52	2.53	5.34	1.9e-07
WF_{SUB-PL}	−137.22	14.26	−9.62	<2e-16
DLF	−107.38	12.69	−8.46	1.2e-15
WF_{SUB-PL} * *DLF*	21.66	2.91	7.44	1.1e-12

Adjusted $R^2 = .719$; $F(4, 295) = 193.00$, $p < 2e\text{-}16$

Note. The frequencies were centered before being entered into the linear regression

range, SUBTLEX-UK frequencies accounted for 9.59 % of the variance (*SD* = 5.00), and BNC frequencies for 6.61 % (*SD* = 4.23) of the variance. In the upper part of the frequency range, SUBTLEX-UK frequencies accounted for 29.73 % (*SD* = 6.9) of the variance, and BNC frequencies for 24.59 % (*SD* = 6.26) of the variance. Interestingly, averaged word frequency accounted for slightly more variance than did SUBTLEX-UK in both lower (mean=10.53 %, *SD* = 5.00) and upper (mean=30.25 %, *SD* = 6.58) ranges. The averaged word frequency was also slightly better at predicting RTs for the full set of words (mean=44.04 %, *SD* = 4.74) than were individual frequency measures (SUBTLEX-PL, mean=43.23 %, *SD* = 4.77; BNC, mean=40.34 %, *SD* = 4.66). We compared R^2 values obtained in the simulations using the Welsh *t*-test. Due to the large number of simulations, all reported differences were statistically significant, except for the difference between averaged word frequencies and SUBTLEX-UK frequencies in the upper part of the frequency range.

Conclusions

We presented new word frequency estimates for Polish based on film and television subtitles and, in two lexical decision experiments, validated their usefulness by comparing them with estimates derived from BS–NCP, as well as with compound frequency estimates derived from the two text corpora.

We found a large advantage of SUBTLEX-PL over BS–NCP when words for which estimates given by the two corpora differed most were used as stimuli. In contrast, when we sampled words randomly, the advantage became less pronounced (Experiment 1) or tended to favor the BS–NCP-derived frequencies (Experiment 2).

These results suggest that the relationship between frequency estimates derived from different corpora and human performance in behavioral tasks may be complex. In particular, this shows that the stimulus selection procedure may affect the outcome of a validation experiment. For a comparative study to be informative, it is essential to find an unbiased method of stimulus selection. Although it is reasonable to assume that the more words included in a validation study, the more relevant its results, it has to be taken into account that even selecting words from a megastudy for validation (e.g., Keuleers et al., 2010) may introduce bias and make it easier for one of the corpora to provide good frequency estimates than do other corpora. For instance, if only mono- and disyllabic words are included in a study, the mean frequency may be shifted, relative to the mean in the full lexicon, because of a negative correlation between word frequency and word length. In such a case, a corpus that does better in predicting behavioral measures in higher parts of the frequency range would be favored. Using the BLP data, we failed to replicate the advantage of a written text corpus in the lower frequency range, although we found a similar overall interaction between word frequency measures. Also, the small total amount of explained variance in the range below the median point in this analysis may suggest that mono- and disyllabic words do not represent the lexicon well in that frequency range.

Moreover, it should be considered whether including a full set of words in validation studies is an optimal choice. If a word frequency distribution of a full lexicon were reflected in a stimulus set of a validation study, due to properties of a Zipfian distribution, the vast majority of words would have to be on the low extreme of the possible frequency range, and, because in linear regression all observations contribute equally to the results, R^2 would be determined mostly in the very low part of the frequency distribution. In this case, the results of linear regression would not be very informative for high-frequency words.

In addition to these methodological aspects, we would like to point out that it is also possible that some properties of the lexicon may have contributed to the pattern of results obtained in the present study. It is possible that during word processing, lemma frequency is a source of facilitation that is stronger for low-frequency words than for high-frequency words. As Table 5 shows, in an exploratory analysis, we observed a statistically significant interaction between word frequency and lemma frequency when these two variables and word length were entered into a linear regression as predictors and RTs obtained in Experiment 2 as a dependent variable. It is possible that this extra facilitation for low-frequency words corresponds to slightly higher frequency estimates for low-frequency words in written text corpora than in subtitle corpora. If that were the case, the advantage of the written text corpus, in comparison with the subtitle corpus observed in the low-frequency range, could be incidental, rather than reflecting a real advantage of written-text corpora.

To fully explore these issues, it would be necessary to conduct analyses across different sets of stimuli and for different languages. Lexical decision megastudies (Balota et al., 2007; Keuleers et al., 2010; Keuleers et al., 2011) provide a good opportunity for such analyses

Nevertheless, even with a validation using a limited set of words, the results of the two experiments suggest that both SUBTLEX-PL and BS–NCP are valuable sources of word frequency estimates. In most cases, we would advise researchers to use the averaged compound measure derived from the two corpora whenever possible. At the same time, we do not have enough evidence to strongly suggest the same practice in other languages. It must also be kept in mind that for certain classes of words, one of the corpora may give strongly biased frequency estimates. We have shown that for BS–NCP, a subset of low-frequency words used mostly in formal communication may belong to such a category.

Availability

SUBTLEX-PL frequencies and compound SUBTLEX-PL/BS–NCP frequencies are available for research purposes and can be downloaded in RData and csv formats from http://crr.ugent.be/subtlex-pl. They can also be accessed online using a Web interface. Frequencies for words with contextual diversity above 2 are also available in the xlsx (Microsoft Excel) format.

The whole word frequency data set for individual words is contained in two files. The first file includes all strings found in the text corpus with rich information about their part-of-speech tags. The columns give information about the following:

- spelling
- spellcheck—whether the string was accepted as a correct word by the Aspell spell-checker
- alphabetical—whether the word contains only alphabetical characters
- nchar—number of characters in the string

SUBTLEX–PL frequency measures:

- freq—count of how many times the type appears in the subtitles
- capit.freq—count of how many times the type was capitalized
- cd—percentage of film subtitles in which the type appears
- cd.count—count of film subtitles in which the type appears
- dom.pos—most frequent part of speech assigned to the type
- dom.pos.freq—how many times this part of speech was assigned to the type
- dom.lemma.pos—dominant lemma[5] for the type
- dom.lemma.pos.freq—how many times this lemma was assigned to the type
- dom.lemma.pos.total.freq—total frequency of the most frequent lemma for the type (across all types)
- all.pos—list of all part-of-speech assignments for the type
- all.pos.freq—list of frequencies for all corresponding part-of-speech assignments in all.pos for the type
- all.lemma.pos—list of all lemma assignments for the type
- all.lemma.pos.freq—list of frequencies for corresponding lemmas in all.lemma.pos for the type
- all.lemma.pos.total.freq—total frequencies (across all types) of all corresponding lemmas in all.lemma.pos
- lg.freq—$\log_{10}$ of subtitle word frequency
- lg.mln.freq—$\log_{10}$ of subtitle word frequency per million
- zipf.freq—Zipf scale word frequency
- lg.cd—$\log_{10}$ of contextual diversity

Compound frequency measures:

- freq.sn.sum—sum of SUBTLEX-PL and BS–NCP word frequencies
- zipf.freq.sn.sum—normalized (Zipf scale) sum of SUBTLEX-PL and BS–NCP word frequencies
- avg.zipf.freq.sn—averaged Zipf frequencies in SUBTLEX-PL and BS–NCP

The second file contains detailed information about lemma frequencies and particular forms for which this lemma was assigned. The columns in this file are the following:

- lemma—spelling of a base form of a lemma
- pos—part-of-speech tag assigned to a lemma
- spelling—word form assigned to a lemma
- freq—total frequency of a lemma or its inflected form
- cd.count—count of unique film subtitles in which the lemma or its inflected form appears
- cd—percentage of unique film subtitles in which the lemma or one of its inflected forms appears

Frequencies for word bigrams are included in a third file giving information about bigram frequency, contextual diversity, and all punctuation marks separating the words and their frequencies.

Acknowledgments This study was supported by an Odysseus grant awarded by the Government of Flanders to M.B. and a subsidy from the Foundation for Polish Science (FOCUS program) awarded to Z.W. We thank Jon Andoni Duñabeitia, Gregory Francis, and an anonymous reviewer for insightful comments on an earlier draft of the manuscript, Adam Przepiórkowski for providing access to the BS–NCP word frequencies, and Jakub Szewczyk for his help with syllabification of Polish words.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814–823. doi:10.1111/j.1467-9280.2006.01787.x

[5] For practical reasons, we assume that lemma is equivalent to a concatenation of a base form of a word and an associated part of speech tag.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67–82. doi:10.1016/j.jml.2009.09.005

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*(2), 283–316. doi:10.1037/0096-3445.133.2.283

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*(3), 445–459. Retrieved from http://link.springer.com/article/10.3758/BF03193014

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

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie), 58*(5), 412–424. doi:10.1027/1618-3169/a000123

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

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods, 45*(2), 422–430. doi:10.3758/s13428-012-0270-5

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One, 5*(6), e10729. Retrieved from doi:10.1371/journal.pone.0010729

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE, 8*(3), e57410. doi:10.1371/journal.pone.0057410

Cuetos Vega, F., González Nosti, M., Barbón Gutiérrez, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica: Revista de metodología y psicología experimental, 32*(2), 133–143. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=3663992

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-Based Word Frequencies as the Best Estimate of Reading Behavior: The Case of Greek. *Frontiers in Psychology*, *1*. doi:10.3389/fpsyg.2010.00218

Gale, W., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics, 2,* 217–237. Retrieved from http://www.grsampson.net/AGtf.html

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal., 52*(12), 5186–5201. doi:10.1016/j.csda.2007.11.008

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627–633. doi:10.3758/BRM.42.3.627

Keuleers, E., Brysbaert, M., & New, B. (2010a). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650. doi:10.3758/BRM.42.3.643

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*. doi:10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*(1), 287–304. doi:10.3758/s13428-011-0118-4

Kilgarriff, A. (2006). BNC database and word frequency lists. Retrieved May 25, 2014, from http://www.kilgarriff.co.uk/bnc-readme.html

Korpus Języka Polskiego Wydawnictwa Naukowego PWN. (n.d.). Retrieved January 9, 2014, from http://korpus.pwn.pl/

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., & Woroniczak, J. (1990). *Słownik frekwencyjny poszczyzny współczesnej*. Kraków: Instytut Języka Polskiego PAN.

Lewis, M. P., Simons, G., & Fennig, C.D. (Eds.). (2013). *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(04). doi:10.1017/S014271640707035X

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly, 11*(1–2), 151–167.

Przepiórkowski, A. (2012). *Narodowy Korpus Jezyka Polskiego: praca zbiorowa*. Warszawa: Wydawnictwo Naukowe PWN.

Przepiórkowski, A., & Instytut Podstaw Informatyki. (2004). *The IPI PAN corpus: preliminary version*. Warszawa: IPI PAN.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37*(1), 118–139. doi:10.1006/jmla.1997.2510

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 776–784. doi:10.1037/a0022531

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295.

Woliński, M. (2006). Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In M. Kłopotek, S. Wierzchoń, & K. Trojanowski (Eds.), *Intelligent Information Processing and Web Mining* (Vol. 35, pp. 511–520). Springer Berlin Heidelberg. Retrieved from doi:10.1007/3-540-33521-8_55

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, 0*(ja), 1–36. doi:10.1080/17470218.2013.850521

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*(2), 307–333.