

# Worldlex: Twitter and blog word frequencies for 66 languages

Manuel Gimenes<sup>1</sup> · Boris New<sup>2</sup>

Published online: 14 July 2015  
© Psychonomic Society, Inc. 2015

**Abstract** Lexical frequency is one of the strongest predictors of word processing time. The frequencies are often calculated from book-based corpora, or more recently from subtitle-based corpora. We present new frequencies based on Twitter, blog posts, or newspapers for 66 languages. We show that these frequencies predict lexical decision reaction times similar to the already existing frequencies, or even better than them. These new frequencies are freely available and may be downloaded from <http://worldlex.lexique.org>.

**Keywords** Word frequency · Cross-language frequency · Twitter · Blogs

The number of occurrences of a word within a corpus is one of the best predictor of word processing time (Howes & Solomon, 1951). High-frequency words are processed more accurately and more rapidly than low-frequency words, both in comprehension and in production (Baayen, Feldman, & Schreuder, 2006; Balota, Cortese, Sergeant-Marshall, Spieler, & Yap, 2004; Monsell, 1991; Yap & Balota, 2009). This word frequency effect was observed in different tasks such as lexical decision tasks (Andrews & Heathcote, 2001; Balota et al., 2004), perceptual identification tasks (Grainger & Jacobs, 1996; Howes & Solomon, 1951), pronunciation tasks (Balota & Chumbley, 1985; Forster & Chambers, 1973), and semantic categorization

tasks (Andrews & Heathcote, 2001; Taft & van Graan, 1998). This word frequency effect is a robust effect, since it was found in many languages.

For a long time, corpora were compiled from written texts, principally books (Kučera & Francis, 1967; Thorndike & Lorge, 1944). Book-based corpora were created in different languages: Brulex (Content, Mousty, & Radeau, 1990) and Frantext, used in the Lexique database in French (New, Pallier, Ferrand, & Matos, 2001); Celex in English, German, and Dutch (Baayen, Piepenbrock, & van Rijn, 1993); and Kučera & Francis in English (Kučera & Francis, 1967). Some corpora were used despite their age and despite some critics (see Brysbaert & New, 2009, for a discussion of the Kučera & Francis corpus).

More recently, another source of corpora was found to be reliable: movie subtitles. The subtitle-based frequencies were first computed in French by New, Brysbaert, Véronis, and Pallier (2007). The authors showed two main results. First, they showed that the subtitle-based frequencies were a better predictor of reaction times than the book-based frequencies. Second, the subtitle-based frequencies were complementary to book-based frequencies. For instance, typical words from spoken language in everyday life were much more frequent in the subtitle-based than in the book-based corpora. Because the book-based and subtitle-based frequencies were shown to be complementary in the analyses (they explained more variance together than separately), the authors concluded that book-based frequencies could be good estimates of written language and that subtitle-based frequencies could be good estimates of spoken language. The subtitle-based frequencies were then created in other languages in which these results have been replicated, such as English (Brysbaert & New, 2009), Dutch (Keuleers, Brysbaert, & New, 2010), Chinese (Cai & Brysbaert, 2010), Greek (Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010), Spanish (Cuetos, Glez-Nosti,

---

✉ Manuel Gimenes  
[manuel.gimenes@univ-poitiers.fr](mailto:manuel.gimenes@univ-poitiers.fr)

<sup>1</sup> Université de Poitiers, CeRCA, CNRS, UMR 7295, MSHS, Poitiers, France

<sup>2</sup> Université de Savoie Mont Blanc, LPNC, S-73000, Chambéry, France

Barbon, & Brysbaert, 2011), German (Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, & Böhl, 2011), British (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), and Polish (Mandera, Keuleers, Wodniecka, & Brysbaert, 2015).

Another source that has yielded good frequency measures was the Internet. The Internet presents two advantages: It is easier to get a large corpus from the Internet than from books (since there is no need to scan documents). Second, the language used on the Internet is more varied than the language in books. Lund and Burgess (1996) proposed a corpus (named HAL) based on approximately 160 million words taken from Usenet newsgroups. Burgess and Livesay (1998) found that the word frequencies from HAL were a better predictor of lexical decision times than the Kučera and Francis (1967) frequencies. Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) reached the same conclusion, and they recommended the HAL frequencies (Balota et al., 2007). According to Balota et al. (2004), the poor performance of the Kučera and Francis frequencies was largely due to the small size of the corpus. In order to investigate this question of the importance of the size of the corpus, Brysbaert and New (2009) selected various sections of the British National Corpus (Leech, Rayson, & Wilson, 2001). The sections were of different sizes (from 0.5 million words to 88 million words). They then correlated the word frequencies in the different sections with the reaction times in the English Lexicon Project (Balota et al., 2007). They showed that the percentage of variance accounted for in the lexical decision times reached its peak when the section size was of 16 million words, especially for low-frequency words. The conclusion was that a corpus of 16 million words seems to be sufficient.

Another important variable in explaining cognitive word processing is contextual diversity (Adelman, Brown, & Quesada, 2006), which is defined as the number of documents in a corpus in which a given word is found. Adelman et al. showed that contextual diversity was a better predictor than word frequency in naming and lexical decision tasks. This result was confirmed with subtitle-based frequencies (Brysbaert & New, 2009).

Nowadays, a great number of languages do not yet have reliable word frequency norms. Moreover, people now spend a lot of time on the Internet.<sup>1</sup> For instance, American people are spending 23 h per week texting. More importantly, this duration has increased since last year, and the proportion of Internet users is increasing in most countries.

The goal of this article is twofold. First, we wanted to make available new word frequencies based on Twitter, blogs, and newspapers for 66 different languages. The distinction between the three sources (Twitter, blogs, newspapers) can be

justified because the language constraints are not the same: The blog and newspaper frequencies are similar because both are not limited in length. The Twitter and blog frequencies are similar because both allow more informal language than newspapers, but they differ in length (a tweet is limited to 140 characters). We also hypothesized that Twitter would be more similar to spoken language, since anybody can produce short text messages that will appear on Twitter, while blogs would be more similar to written language, since it requires more investment to write a blog than to write on Twitter. Second we wanted to test whether these Web frequencies could be as reliable as already known frequencies. In order to test the reliability of these new frequencies, we used reaction times from megastudies available in French, English, Dutch, Malay, and simplified Mandarin Chinese.

## Method

In this section, we describe how the Twitter, blog, and newspaper corpora were collected. Then we describe the word frequencies used from books and subtitles. Finally, we present the five megastudies (French, English, Dutch, Malay, and Chinese) that we used to validate all of these frequencies.

### New frequencies and contextual diversity from Twitter, blogs, and newspapers

The three new frequencies were calculated from a freely available collection of corpora for various languages that was created by Hans Christensen.<sup>2</sup> The documents were collected from public Web pages by a Web crawler, and each entry is tagged with the date of publication. The main characteristics of the three corpora, such as the sizes of the different corpora for the different languages, are presented in Table 1 for the 66 languages. All sources taken together, 51 languages out of the 66 are based on a corpus containing more than 10 million words, and 39 out of the 66 on a corpus containing more than 16 million words. We downloaded the corpora and converted them to lowercase. We then calculated the frequencies of all of the different words in the corpora, and the lists were filtered with the spellchecker Hunspell 1.3.2, or with Aspell 0.60.6.1 (nine languages) if the Hunspell dictionary for that language was not available. The filtering resulting from spellchecking the words was important, because the original lists of words contained a lot of entries having orthographic or typographic errors, as well as foreign language entries. The corpora contain words from foreign languages for two reasons: because the automatic language checker may have confounded two

<sup>1</sup> <http://www.businessnewsdaily.com/4718-weekly-online-social-media-time.html>

<sup>2</sup> <http://www.corpora.heliohost.org>; e-mail adress: hc.corpus@gmail.com

**Table 1** Numbers of words (in millions) and numbers of documents in the three new corpora for each language

Country	Collected Date	Spelling	Blogs		Twitter		News		Total
			NbWords	NbDocs	NbWords	NbDocs	NbWords	NbDocs	
Afrikaans	2011	H	6.2	181,051	2.9	219,559	5.2	152,312	14.3
Albanian	2012	H	11.8	327,840	1.3	102,386	13.1	228,553	26.2
Amharic	2013		1.1	24,579			1.3	32,554	2.4
Arabic	2011–12	H	20	877,403	21	1,641,146	21.7	510,612	62.7
Armenian	2011	A	6	243,097			8.5	156,586	14.6
Azeri	2012	A	5	155,140	0.7	64,838	6.9	140,995	12.7
Bengali	2012		3	105,696			2.9	58,998	5.8
Bosnian	2013		5.6	170,333			6	181,370	11.7
Catalan	2013	H	8.2	187,262	6.5	397,410	5	81,893	19.7
Chinese Simplified	2011		27.9	1,045,472	32.55	1,440,112	33.68	682,472	94.13
Croatian	2012	H	10.6	297,117	4.2	317,257	10.6	227,317	25.5
Czech	2011	H	10.6	293,584	8	565,638	10.8	276,881	29.4
Danish	2010–11	H	29.5	904,546	14.9	1,062,567	27.6	887,016	72
Dutch	2011–12	H	25.6	761,163	21.8	1,671,690	13.9	313,508	61.4
English US	2012	H	38.1	899,288	30.9	2,360,148	35.2	1,010,242	104.2
Estonian	2011	H	13.1	409,501	4.4	388,541	11.9	422,432	29.4
Finnish	2011	A	12.8	439,785	3.2	285,214	10.5	485,758	26.4
French	2011–12	H	35.2	880,655	28.9	2,023,279	20.1	358,001	84.2
Georgian	2011		4.7	181,499			4.8	164,614	9.5
German	2010–11	H	23.4	715,439	24.3	1,936,088	27.1	533,905	74.8
Greek	2011–12	H	19.8	564,281	18.8	1,564,325	18.6	424,397	57.2
Greenlandic	2012						3.7	227,073	3.7
Gujarati	2011	H	5.1	224,047			5	116,482	10.2
Hebrew	2011	H	8.4	269,866	4.7	409,582	8.2	199,047	21.4
Hindi	2011–12	H	6.7	280,267			7	134,268	13.7
Hungarian	2011–13	H	23.3	822,669	19.8	1,819,217	23.8	548,938	66.9
Icelandic	2011	H	8.1	234,021	2.8	230,651	5.7	144,018	16.7
Indonesian	2011–12	H	37.9	1,645,328	39	3,449,770	38.3	1,144,596	115.2
Italian	2011–12	H	26.2	839,919	23.9	1,985,519	29.5	394,465	79.5
Japanese	2011		14.86	664,309	11.91	667,119	14.12	312,916	40.89
Kannada	2011		4	173,154			5	175,824	9.1
Kazakh	2012	H	2.1	77,940			3.5	77,799	5.6
Khmer	2012		2.6	122,528			3.4	55,674	6
Korean	2011		17.6	923,997	18.7	1,572,766	19.4	667,314	55.7
Latvian	2012	H	12.5	374,913	11.3	942,301	12.4	319,428	36.3
Lithuanian	2011	H	4	144,945	1.3	125,387	4.6	149,453	9.9
Macedonian	2012	A	6.4	218,055	2.5	192,612	6.5	144,853	15.4
Malayalam	2011	A	2	102,043	0.3	28,337	1.7	40,484	4
Malaysian	2011	H	8.9	333,607	6.1	611,028	8.9	356,723	23.9
Mongolian	2012	A	4.8	156,390			5.2	108,846	10
Nepali	2013	H	2.5	76,080	0.9	62,726	2.5	54,877	5.9
Norwegian	2011	H	16.9	487,754	12.5	897,939	14.5	554,226	44
Persian	2012	A	4.7	135,767			4	167,898	8.8
Polish	2011–13	H	26.5	852,733	22.5	2,066,716	25.8	698,571	74.8
Portuguese Brazil	2011	H	14.2	600,228	19.5	1,672,477	17	380,983	50.7
Portuguese Europe	2011–12	H	21.5	788,683	22.3	1,799,560	24.2	606,037	68
Punjabi	2012	A	14.8	372,073	12.3	940,256	15.1	306,846	42.2

**Table 1** (continued)

Country	Collected Date	Spelling	Blogs		Twitter		News		Total
			NbWords	NbDocs	NbWords	NbDocs	NbWords	NbDocs	
Romanian	2011–13	H	30.8	834,510	12.7	961,551	31.1	669,306	74.6
Russian	2011–12	H	20.3	753,319	23.1	2,136,329	20.3	456,407	63.6
Serbian (Latin)	2013	H	7.3	212,482	6.4	449,312	7.2	167,587	21
Sinhala	2011	H	5	190,719			5.9	143,970	10.8
Slovak	2011	H	11.2	277,600	1.4	103,163	10	245,660	22.7
Slovenian	2012–13	H	14.1	342,459	6.6	517,244	14.8	255,793	35.5
Spanish South America	2012	H	15.3	570,369	14.5	1,140,487	16	389,620	45.8
Spanish Spain	2011–12	H			29.6	2,136,625	16	584,340	45.6
Swahili	2012	H	5.3	170,168	1.1	123,601	7	228,011	13.4
Swedish	2011–12	H	26.9	774,117	23.2	1,770,655	23.9	855,034	74.1
Tagalog	2012		5.1	184,580	4.6	505,743	4.2	136,199	13.9
Tamil	2011–12		4	205,510	3.6	375,178	3.2	133,706	10.8
Telugu	2011–12	H	5	216,574			4.8	119,265	9.9
Turkish	2011–12	A	22	914,741	20.6	1,924,915	21.1	697,728	63.7
Ukrainian	2011	H	10.8	379,212	6.7	570,684	11.3	306,617	28.8
Urdu	2012		3.3	76,439	0.6	89,109	3.9	74,878	7.7
Uzbek	2012						5.1	148,161	5.1
Vietnamese	2012	H	16.4	402,515	12.2	838,067	17.6	284,419	46.1
Welsh	2013	A	2	41,092			1.8	63,602	3.8

Spelling = Spellchecker used to filter the language's corpus: H, Hunspell; A, Aspell

similar languages, or because parts of a text mainly in one language could be in a different language. Twelve of the languages could not be filtered by a spellchecker because reliable spellcheckers could not be found for them. Because Chinese and Japanese do not separate words with spaces, we tokenized Chinese using the Stanford Word Segmenter 3.5.1, and Japanese with Kuromoji 0.7.7.

#### Frequencies already existing (which we name the “classic frequencies” from now on)

**French** We used the book-based frequency (Frantext) and the subtitle-based frequency (Subtlex-FR) from Lexique 3.80 (New, Pallier, Brysbaert, & Ferrand, 2004; website: [www.lexique.org](http://www.lexique.org)).

**English** We used the HAL frequencies (Burgess & Livesay, 1998) and the Subtlex-US frequencies (Brysbaert & New, 2009).

**Dutch** We used the Celex frequencies (Baayen et al., 1993) and the Subtlex-NL frequencies (Keuleers et al., 2010).

**Malay** We used frequency counts from two corpora based, respectively, on a Malaysian newspaper and a Singaporean newspaper (Yap, Rickard Liow, Jalil, & Faizal, 2010).

**Chinese** Because the Chinese Lexicon Project (see the section about megastudies below) makes available lexical decision reaction times for Chinese simple characters, and because Chinese characters can be isolated words or may be combined with other characters into words, we used two different frequencies: the word frequencies (only the occurrences of the character when it was presented as a single word) and the character frequencies, which included all occurrences of the character (Cai & Brysbaert, 2010).

#### Megastudies in French, English, Dutch, Malay, and Chinese

In order to validate the new frequencies empirically, we ran analyses on megastudies, which are large databases of descriptive and behavioral data. The first such study was published by Balota et al. (2007). In the English Lexicon Project (ELP), they collected naming times and lexical decision times for over 40,000 English words from several hundred participants. Here we tested our new frequencies on English behavioral data coming from the ELP, on French behavioral data (the French Lexicon Project, or FLP; Ferrand et al., 2010), on Dutch (Keuleers, Diependaele, & Brysbaert, 2010), on Malay (the Malay Lexicon Project, or MLP; Yap et al., 2010), and on Chinese (Sze, Rickard Liow, & Yap, 2014).

## Results

### Classic frequencies versus new frequencies

We conducted linear regression analyses to predict standardized lexical decision times according to the classic frequencies, the new frequencies, and all frequencies together. The frequency measures used were the  $\log_{10}$ -transformed raw frequencies (Baayen et al., 2006), on which we added 1 and we used a polynome of degree 3 because Balota et al. (2004) showed a nonlinear relationship between log frequency and lexical decision times. For each regression analysis, we included the number of letters, the number of syllables, and an orthographic neighborhood measure (OLD20; Yarkoni, Balota, & Yap, 2008) as control variables. The number of letters was transformed with a polynome of degree 2 because New, Ferrand, Pallier, and Brysbaert (2006) showed that the relationship between lexical decision times and word length was not linear. All words with an error rate greater than 33 % were excluded from the analyses. Finally, we conducted our analyses on 35,658 words in French, 32,088 words in English, 11,855 words in Dutch, 1,363 words in Malay, and 2,277 words in Chinese.

Table 2 shows the percentages of variance in reaction times explained by the different frequency measures. For each analysis, the value presented in Table 2 is the adjusted  $R^2$  multiplied by 100 (to express it as a percentage). In addition to the adjusted  $R^2$  values, we also

used inferential tests to compare the different linear models. More precisely, when two nested models were compared, we used an analysis of variance test. However, when two nonnested models were compared, we used Vuong's test (Vuong, 1989).

In French, the first line in Table 2 shows the results of the regression analysis when the classic frequencies were used as predictors. The second line presents the results when the three new frequencies were used, and the third line presents the results when all five frequencies were used. The first result to note is that the three new frequencies explained more variance (48.46 %) than the classic frequencies (47.56 %), a difference which was significant ( $R^2$  change = 0.9,  $p < .0001$ ). The second result is that the greatest percentage of variation was explained when all five frequencies were entered as predictors in the regression model (50.27 %): This model performed significantly better than those based on the two classic frequencies ( $R^2$  change = 2.71,  $p < .0001$ ) and the three new frequencies ( $R^2$  change = 1.81,  $p < .0001$ ).

In English, we observed a similar pattern to the one we observed in French: Our new frequencies explained as much variance (68.06 %) as the classic frequencies (67.99 %), since the difference was not significant ( $R^2$  change = 0.07,  $p = .78$ ). The model containing all five of the frequencies explained more variance (69 %) than the models with the new or the classic frequencies alone ( $R^2$  change = 1.01 for the classic frequencies and 0.94 for the new frequencies,  $ps < .0001$ ).

In Dutch, contrary to the result observed in French and in English, the classic frequencies explained more variance (40.2 %) than our new frequencies (38.92 %), and this difference was significant ( $R^2$  change = 1.28,  $p = .004$ ). However, as in French and English, the model containing all five of the frequencies was better, as it explained more variance (42.81 %) than the other two other models ( $R^2$  change = 2.61 for the classic frequencies and 3.89 for the new frequencies,  $ps < .0001$ ). A possible explanation for these slightly worse results from Dutch could be that the language detectors used for collecting the new frequencies often confounded Dutch and Flemish (the Dutch language spoken in Flanders), which would be less the case for the subtitle corpus.

In Malay, we observed that our new frequencies explained a lot more variance (65.87 %) than the classic ones (63.2 %), a difference that was significant ( $R^2$  change = 2.67,  $p < .01$ ). Again, the model including all five frequencies explained more variance (67.44 %) than the other models ( $R^2$  change = 4.24 for the classic frequencies and 1.57 for the new frequencies,  $ps < .0001$ ).

Finally, in Chinese, we ran the analyses using character frequencies and also word frequencies. For the character frequencies, the variance explained by our new frequencies (52.86 %) was greater than that with the classic frequencies (47.56 %), a difference that was significant ( $R^2$  change = 5.3,

**Table 2** Percentages of variance in reaction times (RT) explained by the different frequency measures

Language	Predictors in Regression Models	RT (%)
French (35,658 words)	Frantext + Subtlex-FR	47.56
	Twitter + blogs + newspapers	48.46
	All five frequencies	50.27
English (32,088 words)	HAL + Subtlex-US	67.99
	Twitter + blogs + newspapers	68.06
	All five frequencies	69
Dutch (11,855 words)	Celex + Subtlex-NL	40.2
	Twitter + blogs + newspapers	38.92
	All five frequencies	42.81
Malay (1,363 words)	MalayNews + SingaporNews	63.2
	Twitter + blogs + newspapers	65.87
	All five frequencies	67.44
Chinese Character (2,432 words)	Subtlex-CH	47.56
	Twitter + blogs + newspapers	52.86
	All five frequencies	53.86
Chinese Word (2,277 words)	Subtlex-CH	31.38
	Twitter + blogs + newspapers	35.18
	All five frequencies	37.05

$p < .0001$ ). As in the other languages, the model with all five frequencies explained significantly more variance (53.86 %) than the other models ( $R^2$  change = 6.3 for the classic frequencies and 1.0 for new frequencies,  $ps < .0001$ ). For the word frequencies, we observed the same pattern (all of the differences were significant except for the difference between the new and classic frequencies, which was almost significant). It is worth noting that the variance explained was much greater using the character frequencies than using the word frequencies, suggesting that, as in alphabetic languages, whenever a word is presented the characters in this word are activated (Baayen, Dijkstra, & Schreuder, 1997; New, Brysbaert, Segui, Ferrand, & Rastle, 2004; New & Grainger, 2011). For this reason, the next Chinese analyses in this article will be presented using the character frequencies only.

Because our new frequencies explained either similar amounts of variance or more variance than the classic frequencies for all five of these languages, we can conclude that these new frequencies are reliable alternatives to the frequencies used until now. Furthermore, the fact that the model including all five frequencies explained, in all five languages, more variance than the other models means that these two frequency sources can be complementary. In Malay and in Chinese, our new frequencies gave particularly better results than the classic frequencies. One possible explanation could be that these are the two most recent megastudies. We will come again to this issue in the Discussion.

### Relationship between classic frequencies and new frequencies

Table 3 shows the Pearson correlation coefficients for each pair in the five frequencies (except for Chinese, with only four frequencies). All correlations were significant ( $ps < .0001$ ) in all languages. In French, the new frequency that correlated the strongest with Subtlex-FR was Twitter (.920). The new frequency that correlated the strongest with Frantext was blogs (.972). The same pattern was found in English: Among the three new frequencies, the most correlated with Subtlex-US was Twitter (.925), and the most correlated with HAL was blogs (.988). In Dutch, the new frequency that correlated the strongest with Subtlex-NL was Twitter (.877), and the one that correlated the strongest with Celex was news (.974). In Malay, the new frequency that correlated the strongest with both MINews and SgNews was news; this last result can easily be explained, since all three of these frequencies come from newspaper corpora. In Chinese, the new frequency that correlated the strongest with Subtlex-CH was Twitter (.879). Overall, the results showed that when we want to use a frequency similar to books or HAL, it is better to use blog frequencies. When we want to use a frequency similar to spoken or subtitle frequencies, it is better to use Twitter frequencies.

**Table 3** Correlations for each pair in the five frequencies

		Subtlex-FR	Frantext	Twitter	Blogs	News
French	Subtlex-FR	1.000	.850	<b>.920</b>	.825	.743
	Frantext	.850	1.000	.925	<b>.972</b>	.949
	Twitter	.920	.925	1.000	.948	.901
	Blogs	.825	.972	.948	1.000	.980
	News	.743	.949	.901	.980	1.000
English	Subtlex-US	1.000	.818	<b>.925</b>	.820	.741
	HAL	.818	1.000	.929	<b>.988</b>	.971
	Twitter	.925	.929	1.000	.930	.871
	Blogs	.820	.988	.930	1.000	.975
	News	.741	.971	.871	.975	1.000
Dutch	Subtlex	1.000	.742	<b>.877</b>	.846	.651
	Celex	.742	1.000	.892	.966	<b>.974</b>
	Twitter	.877	.892	1.000	.964	.851
	Blogs	.846	.966	.964	1.000	.925
	News	.651	.974	.851	.925	1.000
Malay	MINews	1.000	.933	.418	.892	<b>.946</b>
	SgNews	.933	1.000	.402	.914	<b>.975</b>
	Twitter	.418	.402	1.000	.615	.434
	Blogs	.892	.914	.615	1.000	.943
	News	.946	.975	.434	.943	1.000
Chinese	Subtlex	1.000	<b>.879</b>	.864	.681	
	Twitter	.879	1.000	.977	.879	
	Blogs	.864	.977	1.000	.916	
	News	.681	.879	.916	1.000	

The best correlation between each classic frequency database and the new frequencies is presented in bold

### Twitter versus blogs versus newspapers

Among the three new frequencies, are all three useful, or could we omit some of them? To answer this question, we ran regression analyses to compare whether the adjusted  $R^2$  decreased when we remove one of the three frequencies. Table 4 presents the percentages of variance in reaction times explained by the new frequency measures. For each analysis, the value presented in Table 4 is the adjusted  $R^2$  multiplied by 100 (to express it as a percentage).

In French, the first line in Table 4 presents the percentage of variance in reaction times explained by the three new frequencies as predictors (48.46 %). The second line presents the regression model with only Twitter and blogs as predictors: The percentage of variance explained is virtually the same (48.4 %). However, because of the important number of observations, this difference was significant ( $p < .0001$ ). When Twitter (48.2 %) or blog (46.19 %) frequencies were removed

from the model (third and fourth lines), we observed further drops in the variance explained ( $p < .0001$ ). In English, we observed a similar pattern: The variances explained were virtually the same with the three frequencies (68.06 %) and when newspaper frequencies were removed (68 %), even though the difference was significant ( $p < .0001$ ). When Twitter (67.18 %) or blog (67.51 %) frequencies were removed, we observed further drops in the variance explained ( $p < .0001$ ). In Dutch, the variances explained were almost the same with the three frequencies (38.92 %) and without the newspaper frequencies (38.86 %), but the difference was significant ( $p = .003$ ). Without Twitter (38.33 %) or without blog frequencies (37.04 %), the drops in variance were greater ( $p < .0001$ ). In Malay, again the variances explained were very similar with all three frequencies (65.87 %) included and with a model based only on blogs and Twitter frequencies (65.91 %) and this difference was not significant ( $p = .64$ ). On the contrary, the variance explained dropped when Twitter (65.3 %) or blog (64.8 %) frequencies were removed from the model ( $p < .0001$ ). Finally, in Chinese, although all differences were significant (all  $p < .01$ ), the difference between the three-frequencies model (52.86 %) and the models without newspapers (52.66 %) or without Twitter (52.53 %) was weak. However, the drop in variance was greater when blogs were removed (51.53 %). Overall, these results indicate that in the five languages, the newspaper frequencies are not crucial if we already use blog and Twitter frequencies.

### Frequency versus contextual diversity

In the book and subtitle corpora, contextual diversity (CD) is a better predictor of word processing than word frequency (Adelman et al., 2006; Brysbaert & New, 2009). We decided to verify whether we could replicate this result with our new corpora. In order to do so, we ran regression analyses. For each of the five new corpora, we compared the word frequency and the CD measure. Table 5 presents the percentages of variance in reaction times explained by the different frequency and CD measures. For each analysis, the value presented in Table 5 is the adjusted  $R^2$  multiplied by 100 (to express it as a percentage).

In French, the first line shows the percentage of variance explained when Twitter frequency or Twitter CD was used as a predictor: The values were extremely similar (43.77 for Twitter frequency and 43.78 for Twitter CD), and the difference was not significant ( $p = .14$ ). The same pattern was observed for blogs (except that the difference was significant:  $p < .0001$ ) and for news ( $p = .4$ ). In English and in Malay, all differences were not significant ( $p > .05$ ). In Dutch, all differences were significant ( $p < .0001$ ), but for each predictor, the frequency was better than the CD. Finally, in Chinese, the variances explained by the frequency measure and the CD measure were very similar, and all of the differences were

nonsignificant. From these results, we can conclude that for the three new corpora (Twitter, blogs, and news), CD is not a better predictor than word frequency.

### Effect of time on the classic and new frequencies

Twitter and blogs are recent tools, as compared to books and even subtitles, since subtitles have been collected from movies released from the beginning of spoken cinema to nowadays. For instance, Twitter was created only in 2006. As a consequence, we predicted that our new frequencies will improve with time and will predict reaction times better and better. A first way to test this prediction is to compare reaction times collected in different years. We compared the evolution of Subtlex, Twitter, and blog frequencies between two megastudies in American English that used the same methodology: the reaction times for young people in Balota et al. (2004) and the reaction times in ELP (Balota et al., 2007). The results are presented in Table 6.

The results show that the variance explained by Subtlex did not improve between 2004 and 2007, contrary to the Twitter ( $+1$  %) and blog ( $+2$  %) frequencies.

### Discussion

The goals of this study were to validate empirically new frequencies derived from Twitter, blog post, and newspaper corpora, and to make available these new frequencies for 66 languages; many of these languages have never had good frequency sources until now.

Our results showed that these new frequencies predict lexical decision reaction times similar to, or even better than, the frequencies that have been used until now, such as book-based and subtitle-based frequencies. This result was found in French, English, Dutch, Malay, and Chinese. Therefore, we can reasonably infer that the new frequencies can be used in other languages, as well. For a great number of languages that have not yet had reliable frequencies, we provide frequencies calculated from Twitter, blog posts, and newspaper corpora. Furthermore, we showed that, for the five languages we analyzed, the newspaper frequencies did not explain much variance that was not already explained by the Twitter or blog frequencies.

Another advantage of these frequencies is that the original corpora are available to download<sup>3</sup> for people looking for extra information, such as the contexts in which the words occur or the frequency of a chain of words.

Surprisingly, in our analyses we did not observe that contextual diversity was a better predictor of latencies than was the word frequency, which is a result that has been replicated

<sup>3</sup> <http://www.corpora.heliohost.org/download.html>

**Table 4** Regression results to compare the three new frequencies

Language	Predictors in Regression Models	RT (%)	$R^2$ Change
French (35,658 words)	Twitter + blogs + news	48.46	
	Twitter + blogs	48.4	0.06
	Blogs + news	48.2	0.26
	Twitter + news	46.19	2.27
English (32,088 words)	Twitter + blogs + news	68.06	
	Twitter + blogs	68	0.06
	Blogs + news	67.18	0.88
	Twitter + news	67.51	0.55
Dutch (11,855 words)	Twitter + blogs + news	38.92	
	Twitter + blogs	38.86	0.06
	Blogs + news	38.33	0.59
	Twitter + news	37.04	1.88
Malay (1,363 words)	Twitter + blogs + news	65.87	
	Twitter + blogs	65.91	−0.04
	Blogs + news	65.03	0.84
	Twitter + news	64.8	1.07
Chinese (2,277 words)	Twitter + blogs + news	52.86	
	Twitter + blogs	52.66	0.2
	Blogs + news	52.53	0.33
	Twitter + news	51.53	1.33

several times (for lexical decision tasks, see, e.g., Keuleers et al., 2010, or Cai & Brysbaert, 2010). A possible explanation could be that the documents in our corpora were too small. For copyright reasons, a document in our corpus contains only a

**Table 5** Regression results to compare the frequency and contextual diversity (CD) measures

Language	Predictors in Regression Models	Frequency	CD	$R^2$ Change
French (35,658 words)	Twitter	43.77	43.78	0.01
	Blogs	48.11	48.16	0.05
	News	44.16	44.17	0.01
English (32,088 words)	Twitter	66.99	66.99	0
	Blogs	66.87	66.88	0.01
	News	64.96	64.95	−0.01
Dutch (11,855 words)	Twitter	35.84	35.81	−0.03
	Blogs	38.22	38.08	−0.14
	News	31.52	31.4	−0.12
Malay (1,363 words)	Twitter	63.17	63.18	0.01
	Blogs	65.04	65.09	0.05
	News	60.69	60.72	0.03
Chinese (2,277 words)	Twitter	50.94	50.9	−0.04
	Blogs	52.19	52.16	−0.03
	News	46.36	46.5	0.14

few sentences (mostly one to three sentences), often one. From this result, we can conclude that when the documents are too small, the differences between frequency and CD disappear, so both indices can be used equally.

Moreover, these new frequencies were collected in identical ways in the different languages, and this can be useful for cross-language studies. Indeed, this will allow researchers to control frequency in very similar ways across possibly very different languages. For example, studies about bilingualism could benefit from Worldlex: If bilinguals have to read the same words in two languages, Worldlex would be a useful tool to control the frequencies across the two languages.

Although the Twitter and blog frequencies are complementary, it can be noted that the blog frequencies were always as good as or better than the Twitter frequencies. A possible explanation would be that, because tweets are often messages typed into a cell phone, the Twitter frequencies would be particularly affected by word length, whereas this bias would not be present for the blog frequencies. To test this prediction, we compared Twitter and blog frequencies for different word lengths: We did not find any argument in favor of this prediction (at least for all words with less than 12 letters). To summarize, for a given language, if blog frequencies are better than Twitter frequencies, this difference occurs whatever a word's length is.

Finally, the sources of these new corpora are very recent (especially Twitter and blogs), and despite their young age, the lexical frequencies are already very good for predicting reaction times in lexical decision tasks. If Twitter and blogs continue to increase in popularity, it is possible that these frequencies will become better and better. Indeed, our analysis comparing English megastudies in 2004 and 2007 suggested that our new frequencies improve with time. This interpretation could also explain why the new frequencies predict reaction times much better than the classic frequencies in Malay and in Chinese, since the two corresponding megastudies are the most recent ones.

### Availability

The WorldLex Twitter, blog, and newspaper frequencies are available at <http://worldlex.lexique.org>. For each language, a word frequency file was created that contains the following information: the raw frequency, the frequency per million

**Table 6** Regression results to compare Subtlex, Twitter, and blog frequencies in 2004 and 2007 (2,511 words)

	RT (%) in 2004	RT (%) in 2007
Subtlex	40.82	40.86
Twitter	38.06	39.02
Blogs	40.14	41.95

words, the contextual diversity, and the percentage of contextual diversity. This information is available for the blog, Twitter, and newspaper corpora (except when one of these corpora was not available, of course). For each language, there are two files: one with the frequencies of all of the character strings that were in the corpus, and one with only the strings that were validated by the spellchecker for the given language. As we already mentioned, the advantage of the file filtered by the spellchecker is that it contains many fewer orthographic or typographic errors than the raw-frequency file. It also does not contain words from foreign languages. However, we have also made the file not filtered by the spellchecker available, because it can contain proper names or new words that were removed by the spellchecker. Except for these specific analyses, we advise researchers to use the spellchecked files.

## 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, 814–823. doi:10.1111/j.1467-9280.2006.01787.x
- Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 514–544. doi:10.1037/0278-7393.27.2.514
- Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37, 94–117. doi:10.1006/jmla.1997.2509
- Baayen, R. H., Feldman, L. F., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 55, 290–313. doi:10.1016/j.jml.2006.03.008
- Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX Lexical Database (CD-ROM)*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.
- Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Access and/or production? *Journal of Memory and Language*, 24, 89–106.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133, 283–316. doi:10.1037/0096-3445.133.2.283
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459. doi:10.3758/BF03193014
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58, 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, 977–990. doi:10.3758/BRM.41.4.977
- Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers*, 30, 272–277. doi:10.3758/BF03200655
- Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, 5(e10729), 1–8. doi:10.1371/journal.pone.0010729
- Content, A., Mousty, P., & Radeau, M. (1990). BRULEX: Une base de données informatisée pour le français écrit et parlé. *L'Année Psychologique*, 90, 551–566.
- Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32, 133–143.
- Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behaviour: The case of Greek. *Frontiers in Language Sciences*, 1(218), 1–12. doi:10.3389/fpsyg.2010.00218
- Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavioral Research and Methods*, 42, 488–496. doi:10.3758/BRM.42.2.488
- Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12, 627–635.
- Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, 103, 518–565. doi:10.1037/0033-295X.103.3.518
- Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, 41, 401–410.
- Keuleers, E., Brysbaert, M., & New, B. (2010a). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42, 643–650. doi:10.3758/BRM.42.3.643
- Keuleers, E., Diependaele, K., & Brysbaert, M. (2010b). 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), 1–15. doi:10.3389/fpsyg.2010.00174
- Kučera, H., & Francis, W. N. (1967). *Computational analyses of present-day American English*. Providence: Brown University Press.
- Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.
- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28, 203–208. doi:10.3758/BF03204766
- Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, 47, 471–483. doi:10.3758/s13428-014-0489-4
- Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 148–197). Hove: Erlbaum.
- New, B., Brysbaert, M., Segui, J., Ferrand, L., & Rastle, K. (2004a). The processing of singular and plural nouns in French and English. *Journal of Memory and Language*, 51, 568–585. doi:10.1016/j.jml.2004.06.010
- New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28, 661–677. doi:10.1017/S014217640707035X
- New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, 13, 45–52. doi:10.3758/BF03193811
- New, B., & Grainger, J. (2011). On letter frequency effects. *Acta Psychologica*, 138, 322–328. doi:10.1016/j.actpsy.2011.07.001

- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004b). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36, 516–524. doi:[10.3758/BF03195598](https://doi.org/10.3758/BF03195598)
- New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 101, 447–462. doi:[10.3406/psy.2001.1341](https://doi.org/10.3406/psy.2001.1341)
- Sze, W. P., Rickard Liow, S. J., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, 46, 263–273. doi:[10.3758/s13428-013-0355-9](https://doi.org/10.3758/s13428-013-0355-9)
- Taft, M., & van Graan, F. (1998). Lack of phonological mediation in a semantic judgment task. *Journal of Memory and Language*, 38, 203–224.
- Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Columbia University, Teachers College.
- van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67, 1176–1190. doi:[10.1080/17470218.2013.850521](https://doi.org/10.1080/17470218.2013.850521)
- Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57, 307–333.
- Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory & Language*, 60, 502–529.
- Yap, M. J., Rickard Liow, S. J., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 Words. *Behavior Research Methods*, 42, 992–1003. doi:[10.3758/BRM.42.4.992](https://doi.org/10.3758/BRM.42.4.992)
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15, 971–979. doi:[10.3758/PBR.15.5.971](https://doi.org/10.3758/PBR.15.5.971)