Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–16

DOI: 10.1177/17470218231190315
qjep.sagepub.com
Sage

SUBTLEX-CY: A new word frequency database for Welsh


Walter JB van Heuven[1], Joshua S Payne[2] and Manon W Jones[3]

Abstract

We present SUBTLEX-CY, a new word frequency database created from a 32-million-word corpus of Welsh television subtitles. An experiment comprising a lexical decision task examined SUBTLEX-CY frequency estimates against words with inconsistent frequencies in a much smaller Welsh corpus that is often used by researchers, the *Cronfa Electroneg o'r Gymraeg* (CEG), and three other Welsh word frequency databases. Words were selected that were classified as low frequency (LF) in SUBTLEX-CY and high frequency (HF) in CEG and compared with words that were classified as medium frequency (MF) in both SUBTLEX-CY and CEG. Reaction time analyses showed that HF words in CEG were responded to more slowly compared to MF words, suggesting that SUBTLEX-CY corpus provides a more reliable estimate of Welsh word frequencies. The new Welsh word frequency database that also includes part-of-speech, contextual diversity, and other lexical information is freely available for research purposes on the Open Science Framework repository at https://osf.io/9gkqm/.

Keywords

Welsh; word frequency; visual word recognition

Received: 16 February 2023; revised: 26 June 2023; accepted: 27 June 2023

Introduction

Welsh is a Celtic language spoken by just over 800,000 people (approximately 30% of the total population; annual population survey, 2022). The language is characterised by several interesting and unusual linguistic features—common to other Celtic languages—such as a verb–subject–object syntactic structure, and morphological features, such as initial consonant mutations (e.g., the possessive pronoun his or her, *ei*, triggers a mutation for subsequent nouns with specific initial consonants; Ball & Müller, 2002). Although Welsh orthography is highly transparent, a number of digraphs (e.g., *dd, th, ph*) and a tendency to form compound nouns mean that written Welsh can at first appear rather complex. Given the history and geography of Wales, it is very rare to find monolingual speakers of Welsh, and instead, a number of regions—including North and West Wales—have populations that are highly fluent in both languages, often acquiring both English and Welsh at home or early on in primary school. Thus, Wales has a population of people with varying degrees of Welsh–English bilingualism, ranging from the highly fluent to beginner level, across the full age range. The distinct linguistic features of either language (different syntax, morphology, levels of orthographic transparency, etc.) are ripe for a proliferation of studies in bilingualism, and indeed, this population has enabled large strides in bilingualism research (cf. Kuipers & Thierry, 2010; Martin et al., 2009; Wu & Thierry, 2013). Creating a large database of Welsh words with associated frequency norms is imperative to ensure that research efforts involving the Welsh language can be conducted efficiently and to a high standard.

Most past research involving written Welsh-language stimuli has relied on the *Cronfa Electroneg o Gymraeg* (CEG; Ellis et al., 2001; see for example Egan et al., 2019;

[1]School of Psychology, University of Nottingham, Nottingham, UK
[2]School of Psychology, Wrexham Glyndŵr University, Wrexham, UK
[3]Department of Psychology, Bangor University, Bangor, UK

Corresponding author:
Walter JB van Heuven, School of Psychology, University of Nottingham, University Park, Nottingham NG7 2RD, UK.
Email: walter.vanheuven@nottingham.ac.uk

Gathercole & Thomas, 2009; Grossi et al., 2010, 2012). CEG is a 1-million words Welsh lexical database that contains frequency counts. Words in this database were selected from a range of modern text types and the intention was to create a Welsh parallel of the Kučera and Francis (1967) database for American English and the Lancaster–Oslo–Bergen (LOB) corpus for British English (Johansson et al., 1978). The CEG corpus has been pivotal in producing Welsh-language psycholinguistic research, yet for research on bilingualism, stimulus selection and matching across languages involves a cumbersome process of weighting frequencies by the size of the respective database to account for inherent biases stemming from the vastly larger, and therefore more reliable, English databases, such as SUBTLEX-UK (van Heuven et al., 2014); a solution that is certainly less than ideal. Other, more recent Welsh word databases include *Corpws Cenedlaethol Cymraeg Cyfoes* (CorCenCC; Knight, Morris, Fitzpatrick, et al., 2020), an 11-million-word database that is highly representative of living Welsh language use, with sources including journals, emails, sermons, road signs, and TV programmes. Nevertheless, for psycholinguistic work focusing on word processing times—and in particular, lexical, written language—frequencies based on film and television subtitles remain better predictors of word processing times than frequencies based on a range of other sources (e.g., Brysbaert, Buchmeier, et al., 2011; Brysbaert, Keuleers, & New, 2011; Brysbaert & New, 2009; Cai & Brysbaert, 2010; New et al., 2007). Thus, the SUBTLEX databases—created from film and television subtitles—provide reliable and precise information on frequency and a number of other indices, and are available in a large number of languages (e.g., Dutch, English, French, Greek, Spanish, and Chinese). Here, we present SUBTLEX-CY, a lexical database of 32-million Welsh words collected from subtitles made available by the Welsh medium broadcaster S4C (broadcasts from 1973 to 2019). Subtitles were collected from a broad range of programmes, including children's programmes, news items, and soap operas. The S4C corpus is substantially larger than existing Welsh corpora (see Table 1). Below, we describe how the corpus was created, provide summary statistics, comparisons with other Welsh corpora, and the first validation study of word frequencies from this corpus. We also examine the rate of cognates and false friends between Welsh and English, and loan words from English: an approach that has not been adopted in previous versions of SUBTLEX in other languages, but may prove fruitful for quantifying linguistic overlap in bilingual communities.

Method

Corpus collection

Welsh subtitles from S4C television broadcasts years 1973–2019 were provided by S4C. The television programmes covered a wide range of genres (e.g., drama, soaps, news, children). The archive included both English and Welsh subtitles in European Broadcasting Union Subtitle Data Exchange format (EBU STL). Files were converted to SubRip Subtitle (SRT) using *stl2srt.py.*[1] Next, based on filename codings used by S4C, the initial set of Welsh subtitles was selected. This resulted in a total of 12,505 files.

Text cleaning

A Python script was created to convert the subtitles to text files. Subtitles not only contain spoken conversation but also information for the hard of hearing that describes sounds or things occurring in the scene, such as CNOC AR Y DRWS (knock on the door), NEGES DESTUN (text message), and FFÔN YN CANU (phone ringing). Such non-spoken material in the subtitles is presented using capital letters. Furthermore, meta-information about the subtitles and other non-spoken text and numbers (e.g., 889) is also often found in subtitles. Several S4C subtitles also contained English translations of some of the Welsh words. These translations were presented between parentheses. All non-spoken material and English translations were removed from the subtitles when these were converted to text. To make sure that the resulting text files contained Welsh language and not English, the language of each text file was determined using *lingua-py.*[2] Four text files identified as English were removed.

Part-of-Speech tagging

After converting the subtitles to text and removing English-only text files, 12,488 Welsh text files remained. To be able to calculate word counts based on the role that words play in sentences, the text corpus was processed with a part-of-speech (PoS) tagger that tokenizes the text and assigns a PoS to each token (e.g., noun, verb, punctuation). There are several PoS taggers for Welsh: WNLT2[3] (Welsh Natural Language Toolkit), CyTag,[4] and TagTeg[5] (G. Prys et al., 2020; G. Prys & Watkins, 2022). WNLT2 and CyTag are rule-based PoS taggers, whereas TagTeg is a statistical tagger based on spaCy (Honnibal et al., 2020) and is trained using an annotated corpus. G. Prys and Watkins (2022) tested the accuracy of these PoS taggers using a corpus of 500 Welsh sentences (7,675 tokens). The results showed that TagTeg reached a token accuracy of 92%, which is significantly higher than the other two PoS taggers. Furthermore, unlike the two rule-based PoS taggers, TagTeg can generalise PoS tags to unfamiliar words. Thus, we decided to use TagTeg to PoS-tag the text files. Unfortunately, TagTeg does not provide lemma information, unlike PoS taggers for other languages. Therefore, the lemma of each word form was looked up automatically in Lecsicon Cymraeg Bangor (the Bangor University Welsh-language lexicon, LCB; Watkins et al., 2021).

Table 1. Number of word forms (types) and corpus size (tokens) of Welsh corpora (> 1-million words) and dictionaries.

Welsh corpora	Word forms (types)	Corpus size (tokens)
CC0 corpus (D. Prys et al., 2021; v21.10) https://github.com/techiaith/corpws-CC0 Corpus of 20,000 sentences and over 180,000 tokens, collected from Wikipedia articles, twitter, out of copyright.	17,068	161,954
More than 100,000 machine-translated sentences from the CoVost Facebook corpus. https://github.com/facebookresearch/covost/	43,850	1,078,379
CEG (Ellis et al., 2001) https://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en (500 samples of 2,000 words, post 1970).	37,192	1,079,131
CorCenCC (Knight, Morris, Fitzpatrick, et al., 2020; Knight, Morris, Tovey-Walsh, et al., 2020) https://corcencc.org National Corpus of Contemporary Welsh (written, spoken, and electronic sources).		14,338,149 (~11.2 million words)
Worldlex (Gimenes & New, 2016) https://worldlex.lexique.org Welsh blogs and newspapers.	89,470	3,794,371
Kynulliad3 (Donnelly, 2013b) http://cymraeg.org.uk/kynulliad3 Word frequency list of 360,000 aligned Welsh–English sentences. Sentences are from the proceedings of the third assembly of the National Assembly for Wales (2007–2011).	41,903	9,377,423
Welsh dictionaries / lexicons		
Eurfa (Donnelly, 2013a) http://eurfa.org.uk Free dictionary that includes word forms from Kynulliad3 and other much smaller corpora.	210,577	
Lecsicon Cymraeg Bangor (LCB; Watkins et al., 2021) https://github.com/techiaith/lecsicon-cymraeg-bangor The Bangor Welsh Lexicon. A comprehensive lexicon of Welsh forms with lemma and morphological information (version 22.07).	496,015	

LCB (version 22.07) contains 496,087 Welsh word forms and includes for each word form, the lemma, PoS, and morphological features. The PoS information provided by TagTeg was used by a Python script to find the lemma of each word form by matching the word form and PoS with those entries in LCB. If the lemma could not be found, the word form was converted to lowercase and again a match was tried based on the word form and PoS. If this failed, only the word form (first in its original form and if failed in lower case) was looked up in LCB to find the lemma. Finally, if again no match was found, the lemma was assumed to be the same as the word form.

After PoS tagging the corpus, a database was created of word type, PoS, and lemma triplets and their counts across all subtitles. This database also contained information in which broadcasts the word type occurred to calculate contextual diversity (Adelman et al., 2006). After removing punctuation from this database, 293,315 types (triplets) and 32,489,072 tokens remained. Next, a lemma frequency database was created from this word type database. In total, the subtitles contained 159,128 lemmas. The SUBTLEX-CY database was created from these two databases. Two SUBTLEX-CY databases with word forms were created, one that included all word forms (171,873 types and 32,489,072 tokens), and using similar criteria as used for SUBTLEX-UK (van Heuven et al., 2014), one database without digits and entries that started with digits or other non-alphanumeric characters except a quote (e.g, ‘d, or those containing a hyphen between letters). Furthermore, to exclude typos and nonwords only word forms that occurred in at least two broadcasts were included. The final cleaned SUBTLEX-UK database contains 87,742 types and 32,242,290 tokens and is recommended to be used by researchers in psycholinguistics.

Each of the files provide frequency, contextual diversity (based on the number of broadcasts in which the word occurred), PoS information, lemma information, and information in which dictionary/lexicon each word occurs. For the dictionary check, each word form was checked against Welsh (cy_GB[6]) and English (en_GB and en_US)[7] Hunspell (version 1.7.1, Ooms, 2022) dictionaries, and words in LCB, Eurfa, and CorCenCC. An overview of the

Table 2. Type and token count for each PoS in SUBTLEX-CY.

PoS (tag)	Types	Tokens
Verb (VERB)	20,883	6,350,803
Noun (NOUN)	39,843	5,426,461
Adposition (ADP)	532	4,075,130
Particle (PART)	47	3,477,554
Pronoun (PRON)	177	3,044,928
Determiner (DET)	81	2,607,370
Adjective (ADJ)	6,906	1,743,716
Adverb (ADV)	649	1,461,178
Conjunction (CONJ)	94	1,317,581
Proper noun (PROPN)	15,729	1,204,209
Auxiliary (AUX)	110	743,521
Interjection (INTJ)	308	304,492
Numeral (NUM)	183	267,402
Other (X)	2,084	168,241
Punctuation (PUNCT)	42	43,549
Symbol (SYM)	74	6,155

Table 3. Language information of the word forms in SUBTLEX-CY (type and token counts, and percentages in parentheses).

Language / dictionaries	Types (%)	Tokens (%)
Welsh (cy_GB and/or LCB and/or Eurfa and/or CorCenCC)	39,485 (45.0)	18,801,515 (58.3)
Welsh and English (Welsh and [en_GB and/or en_US])	3,648 (4.2)	12,056,794 (37.4)
English (en_GB and/or en_US)	25,495 (29.1)	1,050,411 (3.3)
Not found in cy_GB, en_GB, en_US Hunspell dictionaries, LCB, CorCenCC, and Eurfa	19,114 (21.8)	333,570 (1.0)
	87,742	32,242,290

LCB: Lecsicon Cymraeg Bangor; CorCenCC: *Corpws Cenedlaethol Cymraeg Cyfoes.*

number of types and tokens for each PoS category is presented in Table 2.

SUBTLEX-CY contains not only Welsh words but also Welsh–English cognates (e.g., ffrind-*friend*, preifat-*private*; including English loan words that are written identical in Welsh and English, e.g., problem, clown) and Welsh–English false friends/interlingual homographs (e.g., plant [*children*], hen [*old*]). Table 3 provides information about the number of Welsh and English words, and words that can be found in Welsh and English dictionaries (cognates and false friends/interlingual homographs) and entries that did not occur in Hunspell dictionaries, LCB, Eurfa, and CorCenCC.

In addition to word frequency (count of how many times it appears in the subtitles), Zipf values were calculated using Equation (1) provided in van Heuven et al.'s (2014) study and added to the cleaned SUBTLEX-CY database

$$\text{Zipf value} = \log10\left(\frac{\text{count}+1}{\text{tokens per million}+\text{types per million}}\right)+3 \quad (1)$$

The total number of tokens in SUBTLEX-CY is 32.242 million and the number of types is 0.088 million. Thus, the resulting Equation (2) was used to calculate the Zipf values for all entries in SUBTLEX-CY

$$\text{Zipf value} = \log10\left(\frac{\text{count}+1}{32.242+0.088}\right)+3 \quad (2)$$

Furthermore, for each word, the orthographic similarity with other words in the final database was calculated in terms of OLD20 (Yarkoni et al., 2008) and neighbourhood density (Coltheart et al., 1977) using the R package strsim,[8] and these measures were also included in the cleaned SUBTLEX-CY database.

SUBTLEX-CY versus other Welsh word frequency databases

The top-25 of the most frequent words in each database (SUBTLEX-CY, CEG, CorCenCC, Worldlex, and Kynulliad3) is presented in Supplemental Material 1—Appendix 1. The top-25 of each database is similar; however, there are some differences. For example, the most frequent word is "yn" (English translation: "in") in CEG, CorCenCC, Worldlex, and Kynulliad3, whereas in SUBTLEX-CY, which is substantially larger than the other databases, the most frequent word is "i" (English translation is "i").

To explore how the word frequencies differ across the databases, words were selected that occur in all five databases. In total, 9,863 words are in all the five databases, and most are Welsh words (N=9,111). The set also contains form-identical Welsh–English words (cognates/false friends, N=731), English words (N=17), and words that could not be found in Welsh and English Hunspell dictionaries (cy_GB, en_GB, en_US) nor in CorCenCC and Eurfa (N=4). Correlations of the Zipf values between databases were high (see Figure 1). In particular, the correlation between CorCenCC and Worldlex was very high (.908), this is likely due to the use of very similar source material (Welsh online material). Correlations between Kynulliad3 and the other frequency databases are low, this is likely also due to differences in source material. Kynulliad3 frequencies are based on written proceedings of the third assembly of the National Assembly for Wales, whereas the other frequency databases are based on either online material or written/spoken material.

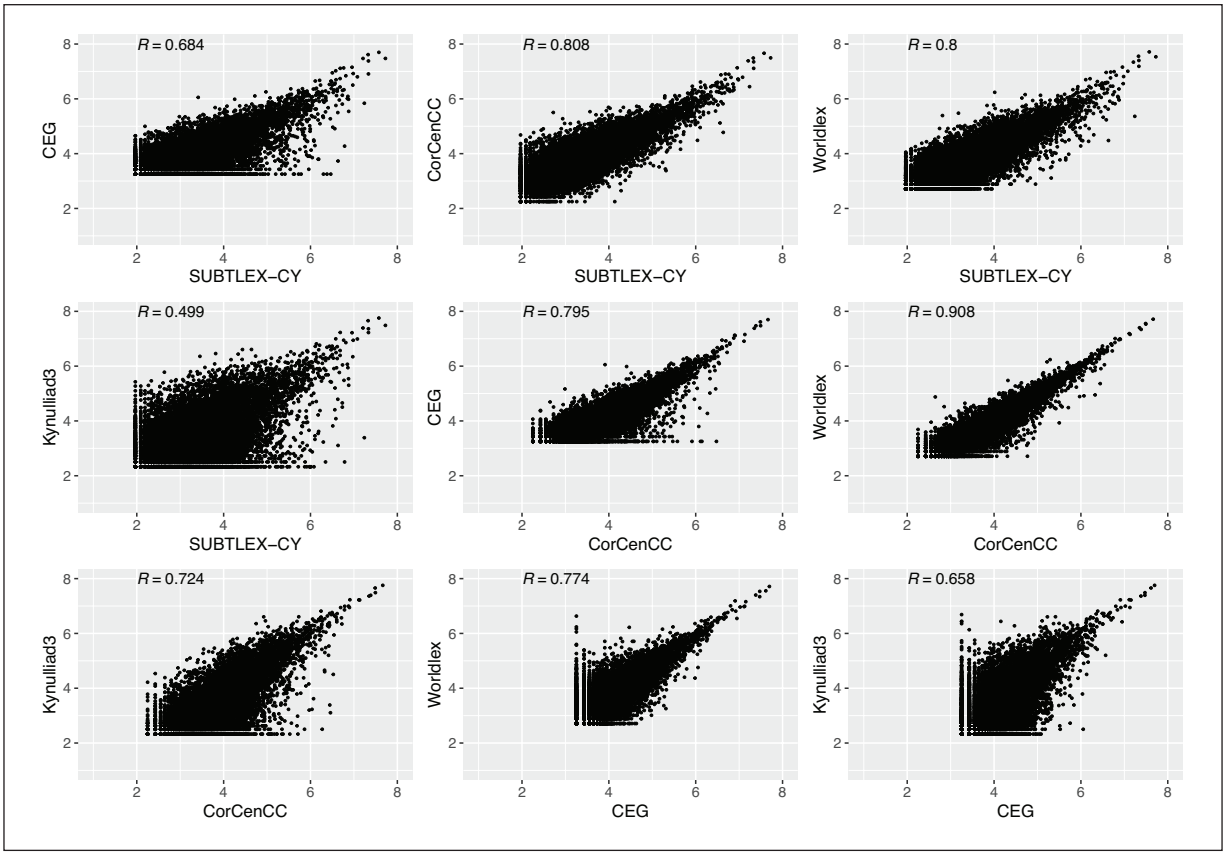

Figure 1. Scatterplot of Zipf values and Pearson correlations (all *ps* < .001) between the five Welsh-word frequency databases (*N* = 9,863).

Next, we explored the frequency distributions of the Welsh words and the Welsh–English form-identical words across the databases (see Figure 2). Interestingly, the Zipf value distributions suggest that the Zipf values for Welsh–English words are higher than for Welsh words (distribution of Welsh–English word more to the right compared to the Welsh word distribution). The mean Zipf values are consistent with this because for all databases, the Zipf values are significantly higher for Welsh–English cognates/false friends than for Welsh words. However, for SUBTLEX-CY, CorCenCC, and Worldlex, the difference in mean Zipf value between Welsh–English words and Welsh words is relatively large, SUBTLEX-CY: 4.31 vs 3.47, $t(824.84) = 22.53$, $p < .0001$; CorCenCC: 4.28 vs 3.85, $t(806.6) = 12.25$, $p < .0001$; and Worldlex: 4.27 vs 3.85, $t(810.05) = 12.22$, $p < .0001$, whereas the differences in CEG and Kynulliad3 are only 0.16 and 0.10, CEG: 4.31 vs 4.15, $t(788.59) = 4.88$, $p < .0001$; Kynulliad3: 3.61 vs 3.51, $t(808.63) = 2.16$, $p = .0308$

Welsh–English form words in SUBTLEX-CY

SUBTLEX-CY contains 3,648 words classified as Welsh and English (cognates/false friends). In total, 3,323 of these words also occur in SUBTLEX-UK (van Heuven et al., 2014) and the correlation between the Zipf values in SUBTLEX-CY and SUBTLEX-UK is moderate ($r = .438$, $p < .001$). However, many of the words are proper names. After removing the proper names based on the PoS tagger information in SUBTLEX-UK, the set of words was reduced to 1,942. The correlation between the Zipf values of these 1,942 cognates/false friends in SUBTLEX-CY and SUBTLEX-UK is higher ($r = .562$). Supplemental Material 1—Appendix 4 shows the top-50 most frequent form-identical cognates/false friends.

Next, we examined the potential disparity between word frequencies in SUBLTEX-CY and in the currently most-used corpus in Welsh-language research, CEG (Ellis et al., 2001). Figure 3 shows words identified as having consistent (mid-range) frequencies in either corpus, and words that are inconsistent across the two corpora; high frequency (HF) in one and low frequency (LF) in the other (see "Method" section of the experiment below for the information of how these words were selected). These classifications were then plotted also for the other Welsh databases (CorCenCC, Worldlex, and Kynulliad3). The resulting pattern of mean frequencies shows a fundamental inconsistency for words identified as LF in SUBTLEX-CY compared to the frequency of these words in the other corpora.

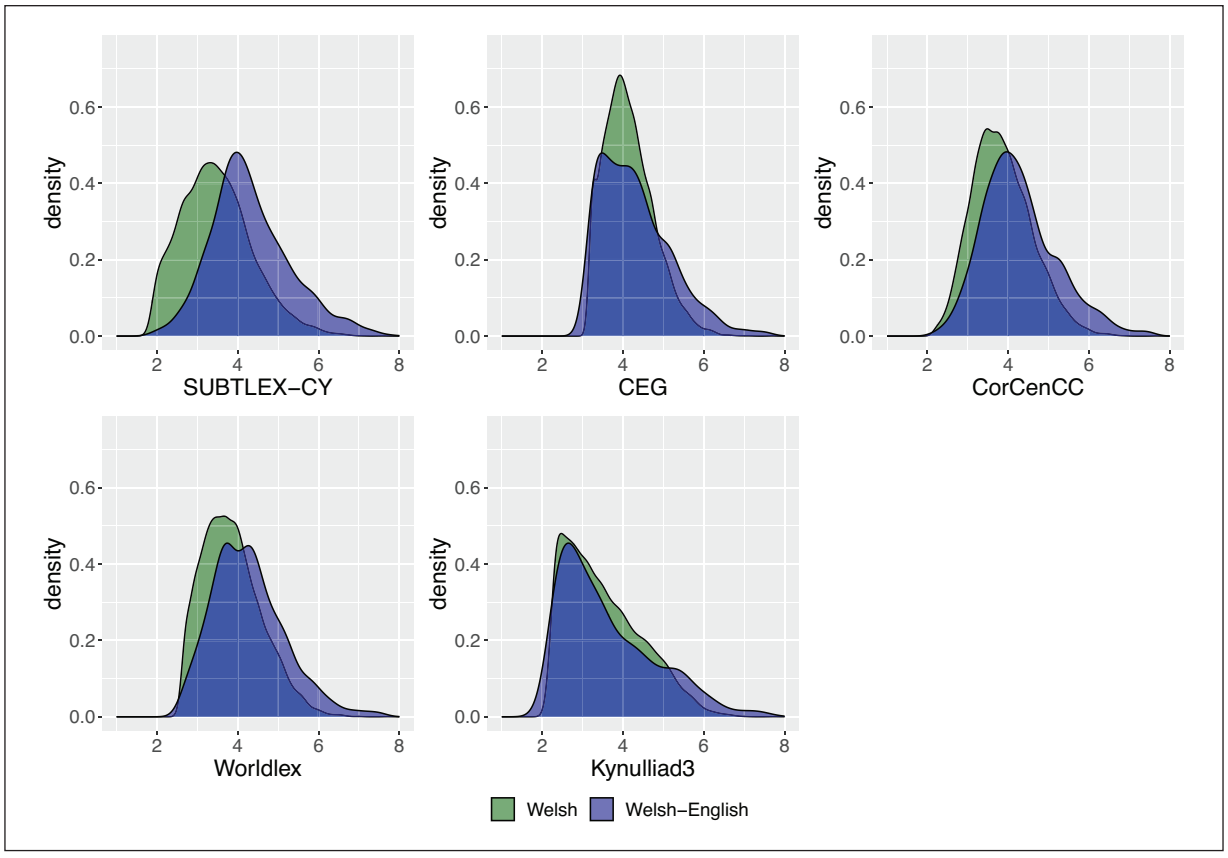

Figure 2. Density plots of Zipf values by language for Welsh words ($N = 9{,}111$) and form-identical Welsh–English words ($N = 731$) that occur in all five databases.

Given this interesting inconsistency, we next conducted an experiment with human participants to assess the fit of SUBTLEX-CY and CEG word frequencies with participant's response times in a Welsh lexical decision task.

Experiment

Method

Participants. Overall, 67 participants completed the experiment, recruited via social media platforms and Prolific.co. Eligibility was restricted to participants aged between 18 and 40 years, who reported oral and written fluency in Welsh and English and reported no language-related disorders. Participants were paid £3.80 for a maximum of 30 min of their time.

Following the initial screening, a language history questionnaire was administered, which required participants to self-rate proficiency in reading, writing, speaking, and comprehension in Welsh, on a scale from 0 to 10. An aggregate self-rated proficiency score is calculated as a mean over the four variables. Despite pre-screening, there was a wide range of self-reported proficiency values in Welsh. Aggregate Welsh proficiency scores ranged from 2.5 to 10 with median of 8.5 ($MAD = 2.22$, $IQR = 3$). Participants who reported self-rated aggregate proficiency of less than 7 were excluded from analyses ($n_{excl} = 23$) as they were not considered proficient enough in Welsh for the purpose of validating a new lexical frequency database. One additional participant with a Welsh proficiency rating of 9.75 was excluded because their accuracy on word trials was 1.11%.

The remaining 43 participants had a mean age of 27.35 ($SD = 7.13$, 18–41) years. In total, 25 participants identified as women, 16 as men, and 2 did not wish to say, with majority right-handed ($n_{right} = 35$, $n_{left} = 5$, $n_{ambi} = 3$). Most participants learned English before the age of 4 years ($n = 37$). However, 51% ($n = 22$) of participants used Welsh and English at home, 28% ($n = 12$) used English only, 16% used Welsh only ($n = 7$), while one participant reported using English and German at home, and another reported use of English, Welsh, and Spanish. Median Welsh proficiency ratings were 8.75 ($MAD = 1.48$). Ethical approval was granted by Wrexham Glyndŵr University.

Stimuli. A total of 25,182 word forms were common to the CEG (Ellis et al., 2001) and a non-final version of SUBTLEX-CY (June 2021). To evaluate the frequency estimates from the SUBTLEX-CY corpus, words with Zipf values greater than or equal to 3.0 were selected to ensure that selected stimuli would be known to most Welsh speakers. The set was further restricted to include stimuli that

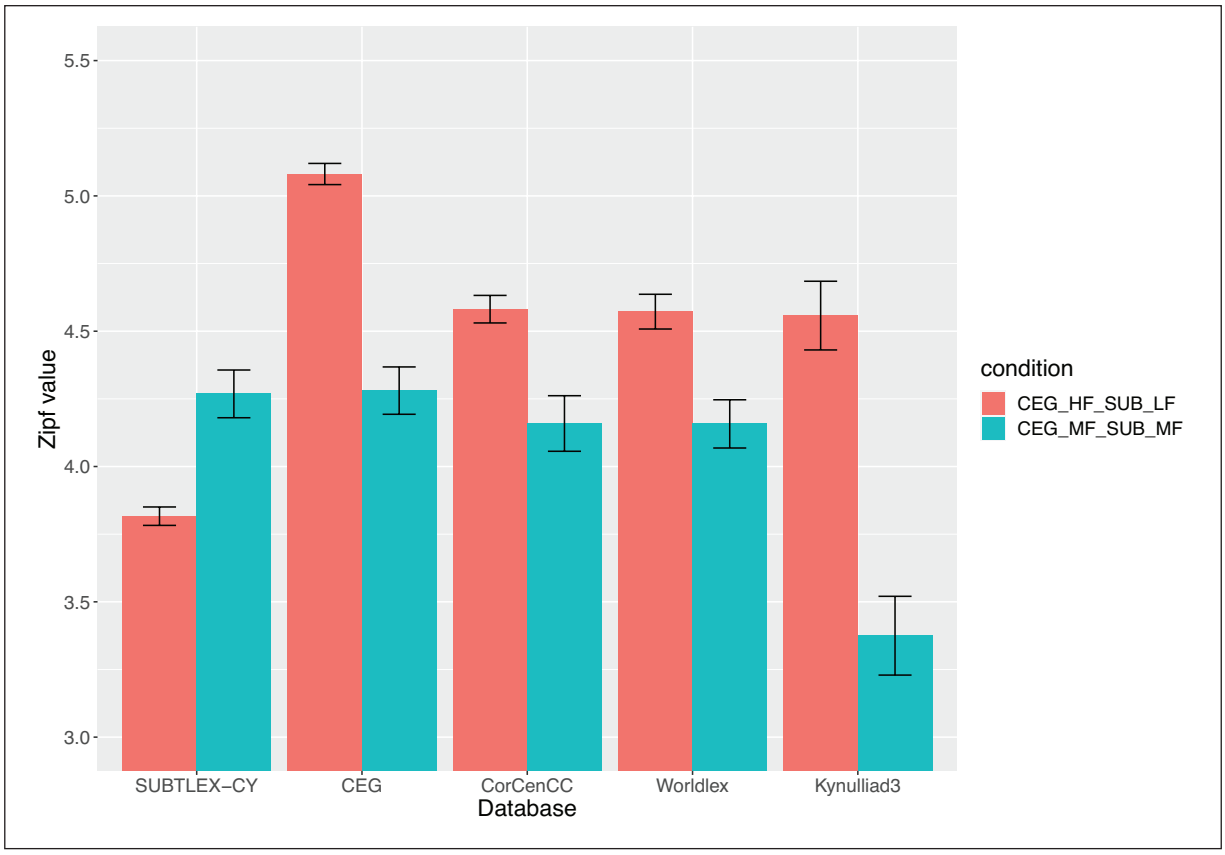

Figure 3. Mean Zipf values and standard error for words in the two conditions in five lexical databases (SUBTLEX-CY, CEG, CorCenCC, Worldlex, and Kynulliad3).

differed in Zipf value estimates between corpora by a minimum of 0.1 and a maximum of 1.1. Words were restricted to a minimum of four letters and a maximum of nine. This resulted in three subsets of words corresponding to three tiers of relative between-corpora differences in Zipf estimates:

- LF-HF: words considered low frequency in CEG but high in SUBTLEX-CY
- MF-MF: words considered moderate frequency in CEG and SUBTLEX-CY
- HF-LF: words considered HF in CEG but low in SUBTLEX-CY

Borrowings from English, Welsh mutations, slang terms, and cognates were excluded from this subset to reduce idiosyncratic responding and facilitation in the case of cognates. The selected stimuli were further restricted based on PoS information so only nouns, adjectives, verbs, and adverbs were retained. As a result, there were insufficient candidates in the LF-HF category to proceed ($n = 18$) and this subset was dropped before matching. OLD20 estimates (Yarkoni et al., 2008) were calculated for the remaining possible targets based on each of the corpora, using the OLD20 function in version 0.3 of the vwr package (Keuleers, 2013). The OLD20 values from the SUBTLEX-CY and CEG corpora were very strongly correlated for words ($\rho = .98$).

Two sets of 45 words were extracted from MF-MF and HF-LF subsets matched on length (tolerance = 0) and OLD20 (tolerance = −0.1 to 0.1)[9] using the LexOPS package (Taylor et al., 2020). Novel pseudowords were generated using Pseudo (van Heuven, 2020) with the SUBTLEX-CY corpus as the dictionary file. Candidate pseudowords were restricted to words between four and nine letters, excluding the 90 targets matched by LexOPS. Welsh includes sets of distinct digraphs that correspond to specific phonemes (ch, dd, ff, ng, ll, ph, rh, th). These were replaced by distinct characters in the SUBTLEX-CY dictionary file (e.g.,: = dd) to ensure digraphs were substituted appropriately in Pseudo. The alphabet was restricted to legal consonants and vowels in Welsh (vowels = a, e, i, o, u, w, y; consonants = b, c, ch, d, dd, f, ff, g, h, l, ll, m, n, ng, p, ph, r, rh, s, t, th) and to position-specific bigram and trigram frequencies of 100 or greater, based on the entire SUBTLEX-CY corpus. Pseudowords that matched words present in SUBTLEX-CY and SUBTLEX-UK (van Heuven et al., 2014) were automatically excluded. Novel pseudowords were generated by substituting one letter in a random position based on the input strings, where vowels

Table 4. Stimulus characteristics of target words and pseudowords.

	CEG_HF_SUB_LF		CEG_MF_SUB_MF		Pseudowords	
	(*n* = 45)		(*n* = 45)		(*n* = 90)	
	M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range
Length	6.56 (1.34)	4 to 9	6.56 (1.34)	4 to 9	6.56 (1.33)	4 to 9
Zipf CEG	5.08 (0.26)	4.61 to 5.74	4.28 (0.59)	3.55 to 6.17		
Zipf SUBTLEX-CY	3.81 (0.23)	3.51 to 4.36	4.27 (0.60)	3.55 to 6.20		
Zipf difference	1.28 (0.14)	1.10 to 1.71	0.01 (0.05)	−0.08 to 0.10		
OLD20 CEG	2.29 (0.59)	1.15 to 3.90	2.36 (0.65)	1.35 to 4.25	2.42 (0.59)	1.45 to 4.35
OLD20 SUBTLEX-CY	1.83 (0.48)	1.00 to 3.20	1.82 (0.48)	1.00 to 3.25	1.85 (0.48)	1.00 to 3.35
PoS	*n* (%)		*n*(%)			
Noun	23 (51)		22 (49)			
Verb	11 (24)		19 (42)			
Adjective	9 (20)		4 (9)			
Adverb	2 (4)					

CEG: *Cronfa Electroneg o'r Gymraeg*; HF: high frequency; LF: low frequency; MF: medium frequency.

were replaced by vowels, and consonants with consonants, producing 88,250 novel pseudowords. OLD20 values were calculated based on SUBTLEX-CY, for all pseudowords, using the OLD20 function in the vwr package for matching. The OLD20 values from the CEG were also calculated and correlated very strongly with OLD20 from SUBTLEX-CY (ρ = .90). LexOPS was used to generate a total of 90 pseudowords matched on length (tolerance = 0) and SUBTLEX-CY OLD20 (tolerance = −0.1 to 0.1) with the HF-LF and MF-MF sets (*n* = 45 each). Following manual inspection of pseudowords by author M.W.J., two pseudowords ("bitchnach," "lineline") were manually replaced with pseudowords of the same length and OLD20 values to avoid spuriously long reaction times (RTs) for the legal but unusual items. Complete stimulus set characteristics are presented in Table 4[10] and the stimuli are presented in Supplemental Material 1—Appendix 2.

Procedure. Participants enrolled into the experiment via a link posted on social media or an advertisement hosted on Prolific.co. All participant-facing study information was presented in Welsh. Participants read the study information and gave informed consent, before completing the lexical decision task. Participants then completed the language history and demographics questionnaire, before being fully debriefed. Participants were presented with their participant public ID and asked to make a record, so they could withdraw their consent after submission should they wish to do so up until a specified date. No requests were received. A 90-min time limit was applied to the study in Gorilla.sc, after which time the participants' data were automatically rejected from the study and an incomplete response was returned to Prolific.co. An additional 65 participants began the study but did not complete it.

The lexical decision task was administered in Welsh via the Gorilla platform (Anwyl-Irvine et al., 2020). Four practice trials were followed by six blocks of 30 trials with a self-paced break between blocks. Participants responded (binary choice keypress) as quickly but as accurately as possible whether individually presented letter strings were real words or pseudowords in Welsh. Each trial began with a fixation cross (250 ms) with a 100-ms blank screen presented before and after. The target stimulus was next presented until a keypress response (or time out at 3,000 ms), followed by a 1,000 ms ISI. Between Blocks 2 and 3, and Blocks 4 and 5, a single-trial silhouette naming task of a cat or a dog was administered as a bot check, which required a mouse response rather than a button press.

Data analysis. Analyses were restricted to correct response times for word trials only (N_{trials} = 3,870). However, 11 trials (0.72%) were excluded due to time out errors. Accuracy was high in both word conditions (M_{HFLF} = 0.92, SD = 0.07; M_{MFMF} = 0.91, SD = 0.06) and 320 (8.3%) incorrect trials were excluded, leaving 3,520 trials for analysis. The glmer function from lme4 (v1.1-29) in R 4.1.3 was used to fit generalised linear mixed models (GLMMs) with inverse Gaussian distribution and identity link functions to the data. The inverse Gaussian better captures the non-negative, positive-skew of response times compared to a Gaussian distribution and better reflects a general theoretical assumption that frequency effects are additive in word recognition rather than interactive or multiplicative (Balota et al., 2013; Lo & Andrews, 2015; Yap & Balota, 2007). An initial intercept-only model with a by-participant random intercept was fit to the data. The inverse Gaussian model was a better fit than a Gaussian model using the lmer function (χ^2 = 2,036.3). The addition of a cross-classified random intercept for items

Table 5. Sequential model comparison showing similar overall model fit but greater variance explained for SUBTLEX-CY frequency estimate.

	CEG model			SUBTLEX-CY model		
Model step	AICc	BIC	R^2 marginal	AICc	BIC	R^2 marginal
Gaussian: 1 + (1|Participant)	43,196.16	43,214.23		43,196.16	43,214.23	
Inverse Gaussian: 1 + (1|Participant)	41,371.04	41,389.11		41,371.04	41,389.11	
+ (1|Target)	40,866.46	40,890.55		40,866.46	40,890.55	
+ (0 + Condition|Participant)	40,862.96	40,899.09		40,862.96	40,899.09	
+ Length	40,864.98	40,907.13		40,864.98	40,907.13	
+ OLD20	47,619.01	47,668.34	.041	40,859.65	40,907.81	.126
+ Zipf estimate	40,860.22	40,914.40	.151	40,849.36	40,903.54	.288
+ Condition	40,851.95	40,912.15	.275	40,848.40	40,908.59	.298
+ Condition: Zipf estimate	40,848.72	40,914.92	.323	40,845.77	40,911.97	.338
+ Welsh proficiency	40,847.05	40,919.27	.415	40,844.21	40,916.42	.428

All values extracted from the compare performance() function in the performance package (Lüdecke et al., 2021); Model fit estimates based on refitting models after exclusion of five influential items and four influential participants, so estimates of model fit reported here are different compared with that described in the model fitting summary.
CEG: *Cronfa Electroneg o'r Gymraeg*; AIC: Akaike information criterion; BIC: Bayesian information criterion.

improved model fit, $\chi^2(1)=619.25$, $p<.001$. Adding a random slope of condition within participant also improved model fit, $\chi^2(2)=11.70$ $p=.003$, which reflects a maximal random effects structure for this experimental design (Matuschek et al., 2017). Continuous fixed effects were mean-centred, and simple effect coding (−1, 1) was applied to the categorical fixed effect of condition.

A single common model was fit by entering word length as a single predictor, followed by OLD20 values as control variables (cf. van Heuven et al., 2014). OLD20 estimates were calculated using the same corpus as relevant Zipf estimates. In the next step, Zipf values based on CEG (Ellis et al., 2001) and SUBTLEX-CY were entered as fixed effects in two parallel models to provide an estimate of variance explained by each estimate of word frequency. Condition was added as fixed factor to each model, followed by the frequency × condition interaction relevant for each model. Each fixed effect improved model fit (see Table 5).

Visual inspection of model assumptions was carried out using the check_model function from the performance R package (Lüdecke et al., 2021). Variance inflation factor (VIF) values were consistently below five for all predictors across all model steps. There was some heterogeneity in residuals with some deviation from normality. This may have indicated missing predictors. Self-reported proficiency was added to the model, which improved model fit and marginal R^2 substantially improved in both models but issues with diagnostics were still present.

Influential cases for both items and participants were examined by calculating Cook's distance estimates using the leave-one-out procedure implemented in the influence function from the influence.ME package (v0.9-9; Niewenhuis et al., 2012). As we were examining influential cases in two parallel models with differing predictors, we set conditions for when influential cases were excluded: (a) values of Cook's distance should exceed 4/43 ($4/n_{\text{participants}}$) for items and participants as conservative cut-off to avoid excluding too many items or participants, while balancing power given the modest sample sizes and (b) items and participants must be influential cases in both models and in the same rank position (i.e., most extreme case in both models). First, influential items were iteratively dropped, excluding four items (gweld [*to see, v*]; prfiysgol [*university, n*]; silffoedd [*shelf, n*], adeiladu [*to build, v*], and clywed [*to hear, v*]) until models began to disagree on the rank order of influential items. Four influential participants were excluded. Model fit and variance-explained improved substantially, although overall substantive patterns of fixed effects did not change. VIF remained below five for all predictors, although some heterogeneity and deviation from normal residuals remained. The final models with a total of 39 participants, 85 items, and 3,061 observations are presented in Table 5. The final models were fit with:

```
glmer(reaction_time ~ length + old20 + condition + ceg_zipf
      + condition:ceg_zipf + (1 + condition|participant) +
      (1|item),
      nAGQ = 0, family = inverse.gaussian(link =
      "identity"),
      control=glmerControl(optimizer="bobyqa", optCtrl=list
      (maxfun=2e5))

glmer(reaction_time ~ length + old20 + condition + subtlex_
      zipf + condition:subtlex_zipf + (1 + condition|
      participant) + (1|item),
      nAGQ = 0, family = inverse.gaussian(link =
      "identity"),
      control=glmerControl(optimizer="bobyqa," optCtrl=list
      (maxfun=2e5))
```

Table 6. Final model coefficients..

	CEG		SUBTLEX-CY	
	Estimates	*CI*	Estimates	*CI*
Intercept	708.38	[671.83, 744.93]	823.87	[793.68, 854.07]
Length	−31.97	[−53.54, −10.39]	−37.81	[−59.48, −16.14]
OLD20 estimate	62.12	[12.88, 111.35]	104.89	[38.61, 171.17]
Zipf estimate	−181.21	[−228.26, −134.16]	−168.54	[−215.07, −122.02]
log(CD)	107.11	[57.59, 156.63]	−18.31	[−58.77, 22.14]
Condition	162.16	[70.13, 254.20]	−168.54	[−215.07, −122.02]
Zipf: Condition	−41.1	[−62.52, −19.67]	−41.27	[−62.68, −19.87]
Welsh proficiency	708.38	[671.83, 744.93]	823.87	[793.68, 854.07]
σ^2	0.01		0.01	
τ_{00}	3,476.54 items		3,342.69 items	
	3,774.26 participant		3,769.26 participant	
τ_{11}	1,271.83 participant: condition		1,268.16 participant: condition	
ρ_{01}	−0.16 participant		−0.16 participant	
ICC	1		1	
N	85 items		85 items	
	39 participants		39 participants	
Observations	3,061		3,061	
Marginal R^2/ conditional R^2	.415/1.000		.428/1.000	

CEG: *Cronfa Electroneg o'r Gymraeg*; CI: confidence interval; CD: contextual diversity, ICC: intraclass correlation coefficient.

The summary of model-by-model fit statistics (AICc, BIC, marginal pseudo-R2) in Table 5 is based on the compare_performance function from the performance package (Lüdecke et al., 2021). Final model coefficients and fit statistics are presented in Table 6.

Results

Figure 4 shows predicted response times based on random effect estimates for participants for CEG and SUBTLEX-CY Zipf estimates, separated by condition. In both final models, a weak facilitatory effect of increasing word length was observed. The effect of increased orthographic similarity (lower OLD20 values) on Welsh lexical decisions was facilitatory, just as in English (Yarkoni et al., 2008). The facilitatory effect of OLD20 was stronger in the SUBTLEX-CY model compared to the CEG, producing a better overall fit, even at this stage. Model fit improved greatly when frequency was added to both models. Both models showed strong, monotonic effects of frequency overall. However, the current experiment is based on observations that a subset of items according to CEG Zipf estimates is relatively HF compared to the Zipf estimates from SUBTLEX-CY. If estimates from CEG are reliable, we would expect response times to the HF-LF condition to be *faster* than the MF-MF condition. In contrast, Zipf estimates for the same items in this HF-LF subset are similar to the MF-MF subset, so we would not expect responses to be systematically faster or slower between conditions if SUBTLEX-CY estimates are better. In the case of the CEG model, responses are estimated to be 107 ms slower on average ($SE = 25$) to items in the HF-LF condition compared to the MF-MF condition. In contrast, the SUBTLEX-CY model produced the pattern of effects expected by a reliable lexical corpus—a small difference of −18 ms ($SD = 21$) for HF-LF items compared to MF-MF items. Figure 4 clearly shows response times to the cluster of pink HF-LF items shifted to the right relative to the MF-MF items in green for the CEG model, but much more similar estimates for both categories in the SUBTLEX-CY model. Differences in patterns of fit are also observed at this point in the process. For the CEG model, the addition of Zipf to the model including length and OLD20 improved R^2 from .041 to .151 ($R\Delta = .11$), with addition of condition taking R^2 to .275 ($R\Delta = .124$). In the SUBTLEX-CY model, the addition of the Zipf estimates increased variance accounted for from $R^2 = .162$ to $R^2 = .288$ ($R\Delta = .162$); the addition of condition only increasing R^2 to .298. Both models showed strong interactions between condition and Zipf estimates, but this is largely accounted for by the restriction of range in the HF-LF set of items. When self-reported proficiency was added to both models, fit improved substantially in both models, showing a distinct association between higher self-reported proficiency and faster response times. That overall fit is similar in both final models is not surprising given that all factors measured are accounted for in the two parallel models, using the same outcome data.

In Supplemental Material 2, we report models updated with final Zipf and OLD20 values from SUBTLEX-CY

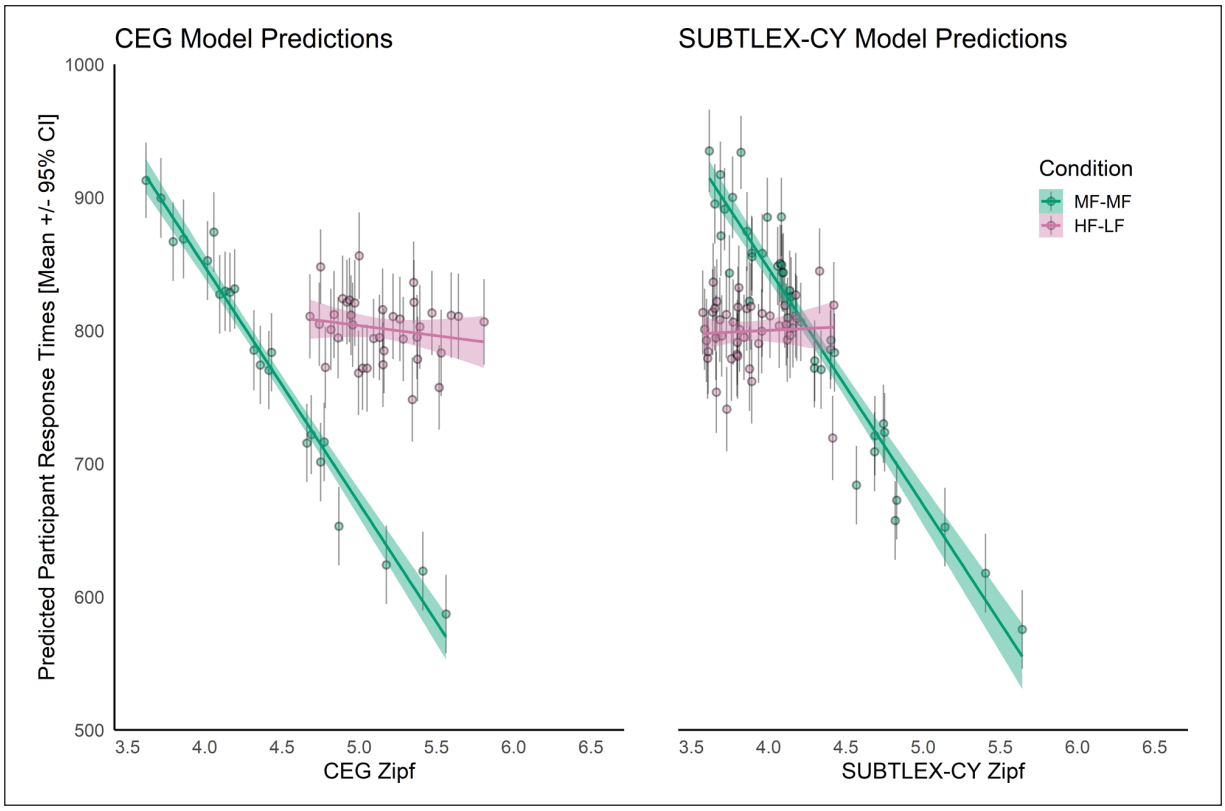

Figure 4. Predicted slopes and predicted by-participant mean (± 95% CI) response times. Items classified as higher frequency by the CEG corpus (HF-LF) showed slower response times compared to moderate frequency items, consistent across corpora. Mean response times in both conditions were similar for the SUBTLEX-CY corpora. Frequency slopes are different between conditions in both models, which may reflect restriction of range.

because the values used in the analyses above were based on a non-final version of SUBTLEX-CY. The means and range of the updated values are reported in Table S1. Importantly, the values are similar to those reported in Table 4. Furthermore, the analyses reported in Supplemental Material 2 (see Tables S2 and S3 and Figure S1) also included OLD20 values of accent-corrected words from CEG because accents were included in the original CEG file with an addition character after the vowel. The models differ to those reported above in that only one influential item was removed as a common case across all three models. The same four influential participants were identified and excluded. Although the Marginal R^2 values in these models are slightly less than in the original models presented above, they can be explained by the retention of more influential items across models. Importantly, the analyses revealed a similar pattern as above.

Discussion

Here, we present SUBTLEX-CY, a new database of Welsh word frequencies based on Welsh television subtitles. We found that SUBTLEX-CY is a more reliable estimator of word frequency compared to CEG and other Welsh word frequency databases. Our experiment, specifically focused on comparing SUBTLEX-CY with CEG because that has been the most commonly used word frequency database. The analyses revealed that lexical variables calculated from SUBTLEX-CY provided better estimates of response times compared to CEG. The amount of variance explained by Zipf estimates was much greater for the SUBTLEX-CY model ($R^2 = .288$) than the CEG model ($R^2 = .151$), where length, OLD20, and Zipf estimates were included. OLD20 estimates from SUBTLEX-CY may also provide a better source of orthographic similarity estimates compared to CEG. Crucially, target words identified as higher frequency in the CEG corpus were actually responded to *more slowly* on average than those of a more moderate frequency. The estimates from the SUBTLEX-CY model showed no such differentiation between stimulus sets. The pattern produced in the CEG model is exactly opposite of what would be expected based of the Zipf values alone (see Figure 4).

Our results demonstrate that television subtitles provide a better estimate of lexical word frequencies, measured here in Zipf values, than other sources, including written (e.g., CEG; Ellis et al., 2001) and spoken, written, and electronic sources (e.g., CorCenCC; Knight, Morris,

Fitzpatrick, et al., 2020). Even though SUBTLEX-CY is based on spoken sources (subtitles reflect the spoken language and subtitles were likely not visible for most viewers of the broadcasts), word frequency estimates are better than those of CorCenCC, Worldlex, and Kynulliad3. The increased reliability of SUBTLEX-CY can be attributed to its size: 32 million words compared with the 1 million available in CEG, its most widely used competitor for psycholinguistic research, and the 11 million words in CorCenCC, 4 million words in Worldlex, and 9 million words in Kynulliad3. Brysbaert and New (2009) showed that a corpus smaller than 16 million words does not provide reliable frequency estimates for LF words (below 10 per million).

Another reason why the frequency estimates in SUBTLEX-CY are better at predicting lexical decision latencies of Welsh speakers is that the estimates are based on spoken material that covers a much wider range of genres (e.g., children's programmes, news programmes, soaps, drama, films, and sport) than the material in other Welsh frequency databases. Furthermore, it reflects everyday spoken Welsh language that has been broadcasted by S4C and likely encountered by many people living in Wales.

The Pearson correlation between item RTs in our experiment and SUBTLEX-CY Zipf values was, however, notably modest (−.474), despite being the strongest correlation overall relative to other corpora (CEG: −.366, CorCenCC: −.454, Worldlex: −.396, and Kynulliad3: −.251). The modest association might be accounted for by a range of uncontrolled participant factors. First, proficiency of participants in this study was variable and limited to simple self-report measures. A more robust assessment of proficiency and other factors, such as language dominance, would be beneficial across larger samples to examine the influence of such factors on the strength of association between frequency and response times. SUBTLEX-CY could be used to develop a rapid and readily available assessment of Welsh proficiency to further this end, similar to LexTALE (Lemhöfer & Broersma, 2012). Second, dialectical idiosyncrasies are quite frequent in Welsh over relatively small geographic areas (Ball & Williams, 2001), but particularly in terms of a North–South divide (Mayr & Davies, 2011). Given the relatively small available sample size for this study, we did not collect broader information on geographic area of language context. Follow-up studies with much larger samples, focusing on further validation of the SUBTLEX-CY frequencies in reading and other language domains, and how patterns vary as a function of contextual factors will be necessary to further evaluate the word frequency estimates of this new Welsh corpus.

The analyses revealed a weak facilitation effect of word length. The length of the Welsh words in the experiment ranged from 4 to 9 letters (mean 6.56 letters), indicating that across this range, there is a slight facilitation effect, which contrasts with New et al.'s (2006) findings of a facilitation effect for 3–5 letter English words. In contrast to English, Welsh orthography is very transparent, which might be the reason for the difference in terms of the effects of word length between these languages.

In the analyses so far, we focused on CEG and SUBTLEX-CY word frequencies. However, contextual diversity (CD) introduced by Adelman et al.'s (2006) study has been found to be a very good predictor of reaction times, often outperforming word frequency (for a recent review, see Caldwell-Harris, 2021). Although CD and word frequency are highly associated, it has been suggested that they reflect different brain mechanisms (Vergara-Martínez et al., 2017). Because of the current debate over the value of word frequency and CD in word recognition (e.g., Brysbaert & New, 2009; Hollis, 2020; Johns, 2021; Johns et al., 2016; Johns & Jones, 2022), we conducted some further analyses with CD, which is also provided in the SUBTLEX-CY database. Correlations revealed that log10(CD) correlated higher with RTs than SUBTLEX-CY Zipf values (−.495 vs −.474). However, as expected, CD and Zipf values are highly correlated (.965) for the stimuli used in the present experiment, and across all words in SUBTLEX-CY (.981). Next, we investigated if a model with log10(CD) instead of word frequency (Zipf values) could explain more variance. Tables S3 and S4 and Figure S1 in Supplemental Material 2 show that the model that includes CD is very similar to the model with Zipf values, in fact the explained variance is the same. This may reflect more recent discussions that contextual diversity offers little over other psycholinguistic factors, such as word burstiness (e.g., Hollis, 2020), or that count-based measures may lack sufficient ecological and semantic richness as a measure of contextual diversity (e.g., Johns & Jones, 2022). However, this study was not designed to assess effects of contextual diversity, and the stimuli were highly restricted by design, making any firm conclusions impossible at this stage.

The TagTeg PoS tagger (G. Prys et al., 2020; G. Prys & Watkins, 2022) was used to obtain PoS information of each word in the subtitles. As mentioned earlier, the accuracy of this PoS tagger is much better than other Welsh PoS taggers (G. Prys & Watkins, 2022). However, the accuracy is lower than PoS taggers available for English, for example, Stanford CoreNL (Manning et al., 2014) and spaCy (Honnibal et al., 2020). Thus, the PoS tag information should be used with caution. Hopefully, a new Welsh PoS tagger with a higher accuracy will become available in the future.

While this article presents a comparison of SUBTLEX-CY with other Welsh corpora and participant behaviour, we also considered the linguistic overlap with English in the form of form-identical cognates and false friends. Over 3,000 words with identical orthography were identified between SUBTLEX-CY and SUBTLEX-UK (van Heuven et al., 2014). After removing proper names, a

total of 1,942 words with identical orthography (cognates and false friends) remained. For these words, the Welsh and English Zipf values showed a moderate correlation, but further work is needed to identify those words in the list that are Welsh–English cognates and those that are false friends. Overall, this corpus offers a resource that can enrich research on bilingual language processing and provides a platform for other foundational psycholinguistic validation studies in Welsh, which until now have been sorely lacking.

The SUBTLEX-CY word frequency database is available for research purposes on the Open Science Framework repository at https://osf.io/9gkqm/. The recommended database is a file with word types that occurred at least in two or more S4C broadcasts. A file with all word types (include numbers) encountered in the PoS tagged subtitles is also available. The files also contain information about contextual diversity in terms of the number of broadcasts in which each word type occurs. Furthermore, a file is available with all 1,942 Welsh–English form-identical words (without proper names) encountered in at least two or more S4C broadcasts and observed in SUBTLEX-CY and SUBTLEX-UK. More details about the content of these files can be found below. Together with these files, materials from the experiment, the R scripts used to analyse the data and R scripts to create the tables and the figures, are available on the Open Science Framework repository.

SUBTLEX-CY files:

1. **SUBTLEX-CY is available as an Excel file (SUBTLEX-CY.xlsx) and as a tab-delimited text file (SUBTLEX-CY.txt).** Both files are identical and have 25 columns and 87,742 rows (excluding the header of the file). They contain word types that occur in at least two S4C broadcasts and that only contain letters (no digits or no word types that start with digits or contain non-alphanumeric symbols). The columns in the files provide the following information:

- Word type in lowercase [Spelling]
- Number of times the word type has been counted in all subtitles [SpellingFreq]
- Length of the word type in number of characters [nchar]
- Zipf value of word type [Zipf]
- OLD20 of the word type [OLD20]
- Orthographic neighbourhood density of the word [ColtheartN]
- The number of broadcasts in which the word type was observed [CD]
- Hunspell Dictionaries (cy_GB, en_GB, en_US) and Welsh corpus/lexicon/dictionary (CorCenCC, Eurfa, LCB) in which the word occurs [Dicts]
- Language (Welsh, English, Welsh–English) of the word [Language]
- All PoS tags associated with the word type [AllPoS]
- All lemmas associated with the word type and PoS tag [AllPoSLemmas]
- All lemmas associated with the word type [AllLemmas]
- Number of times each PoS tag associated with the word has been counted in all subtitles [AllPoSFreq]
- Number of times each lemma associated with the word has been counted in all subtitles [AllLemmaFreq]
- The dominant PoS of the word [DomPoS]
- Number of times the dominant PoS of the word type was observed in all subtitles [DomPoSFreq]
- The number of broadcasts in which the dominant PoS of the word type was observed [DomPoSCD]
- Lemma of the dominant PoS of the word [DomPoSLemma]
- Frequency count of the lemma of the dominant PoS of the word type [DomPoSLemmaFreq]
- All spellings of the word (indicating lower and uppercase characters) [RawWords]
- Frequency counts of the spellings of the word [RawWordsFreq]
- Dominant spelling of the word [DomRawWord]
- Frequency count of the dominant spelling of the word [DomRawWordFreq]

2. **SUBTLEX-CY_all.txt** (21 columns × 171,873 rows) with all word types (including numbers) in the subtitles. The file contains the same columns as the file SUBTLEX-CY.txt, except for the columns: nchar, Zipf, CD, and ColtheartN. An addition column ID is included to indicate the row number.
3. **Welsh–English_words.txt** (4 columns × 1,942 rows) with Welsh–English form-identical cognates/false friends that occur in SUBTLEX-CY and SUBTLEX-UK.

- Spelling of word in lowercase [Spelling]
- Zipf value of the word in SUBTLEX-CY [Zipf.subtlex_cy]
- Zipf value of the word in SUBTLEX-UK [Zipf.subtlex_uk]
- List of Hunspell dictionaries and lexicons in which the word occurs [Dicts]

Acknowledgements

The authors thank the Access Service Coordinators and other staff at S4C who enabled access to their subtitles.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Walter JB Van Heuven https://orcid.org/0000-0003-3183-4449

Data availability statement

Datasets generated during or analysed during the current study are available in the Open Science Framework repository, https://osf.io/9gkqm/. The subtitle files used to create SUBTLEX-CY could not be made available because of copyright reasons. SUBTLEX-CY files are available under the CC BY-NC-ND 4.0 licence.

Supplemental Material

The supplementary material is available at qjep.sagepub.com.

Notes

1. https://github.com/yanncoupin/stl2srt
2. https://github.com/pemistahl/lingua-py (version 1.13)
3. https://sourceforge.net/projects/wnlt-project/
4. https://github.com/CorCenCC/CyTag
5. https://github.com/techiaith/model-tagiwr-spacy-cy
6. https://github.com/techiaith/hunspell-cy (Hunspell Cymraeg 07/2022)
7. https://github.com/marcoagpinto/aoo-mozilla-en-dict (en_GB version 3.1.7: 2023-02-01gb, en_US version 2.91: 2020-12-07us)
8. https://github.com/waltervanheuven/strsim (version 1.2.2)
9. Tolerance refers to the strictness of the LexOPS matching algorithm between pairs of items for a given variable. A tolerance of zero forces the algorithm to search for an exact match in word length between pairs of items, whereas a tolerance between -0.1 to 0.1 would allow slight deviations in OLD20 values between pairs of items.
10. The values presented in Table 4 were based on a non-final version of SUBTLEX-CY and a non-accent-corrected version of CEG. The values based on the final version of SUBTLEX-CY and the accent-corrected version of CEG are presented in Table S1 in Supplemental Material 2.

References

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

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407. https://doi.org/10.3758/s13428-019-01237-x

Ball, M. J., & Müller, N. (2002). The use of the terms phonetics and phonology in the description of disordered speech. *Advances in Speech Language Pathology*, *4*(2), 95–108. https://doi.org/10.1080/14417040210001669321

Ball, M. J., & Williams, B. (2001). *Welsh phonetics*. Edwin Mellen Press.

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: The influence of trial history and data transformations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1563–1571. https://doi.org/10.1037/a0032186

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*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, *2*, Article 27. https://doi.org/10.3389/fpsyg.2011.00027

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

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLOS ONE*, *5*(6), Article e10729. https://doi.org/10.1371/journal.pone.0010729

Caldwell-Harris, C. L. (2021). Frequency effects in reading are powerful—But is contextual diversity the more important variable. *Language and Linguistics Compass*, *15*(12), e12444. https://doi.org/10.1111/lnc3.12444

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance* (*Vol. VI*, pp. 535–555). Academic Press.

Donnelly, K. (2013a). *Eurfa* (v3.0). http://eurfa.org.uk

Donnelly, K. (2013b). *Kynulliad3: A corpus of 360,000 aligned Welsh and English sentences from the Third Assembly (2007-2011) of the National Assembly for Wales*. http://cymraeg.org.uk/kynulliad3

Egan, C., Oppenheim, G. M., Saville, C., Moll, K., & Jones, M. W. (2019). Bilinguals apply language-specific grain sizes during sentence reading. *Cognition*, *193*, 104018. https://doi.org/10.1016/j.cognition.2019.104018

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N. (2001). *Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh*. https://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en

Gathercole, V. C. M., & Thomas, E. M. (2009). Bilingual first-language development: Dominant language takeover, threatened minority language take-up. *Bilingualism: Language and Cognition*, *12*(2), 213–237. https://doi.org/10.1017/S1366728909004015

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*, 963–972. https://doi.org/10.3758/s13428-015-0621-0

Grossi, G., Savill, N., Thomas, E., & Thierry, G. (2010). Posterior N1 asymmetry to English and Welsh words in early and late English-Welsh bilinguals. *Biological Psychology*, *85*(1), 124–133. https://doi.org/10.1016/j.biopsycho.2010.06.003

Grossi, G., Savill, N., Thomas, E., & Thierry, G. (2012). Electrophysiological cross-language neighborhood density effects in late and early English-Welsh bilinguals. *Frontiers in Psychology*, *3*, Article 408. https://doi.org/10.3389/fpsyg.2012.00408

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language*, *114*, 104146. https://doi.org/10.1016/j.jml.2020.104146

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python*. https://doi.org/10.5281/zenodo.1212303

Johansson, S., Leech, G. N., & Goodluck, H. (1978). *The Lancaster-Oslo/Bergen corpus of British English*. Department of English, University of Oslo.

Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review*, *128*(3), 525–557. http://dx.doi.org/10.1037/rev0000265

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, *23*(4), 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

Johns, B. T., & Jones, M. N. (2022). Content matters: Measures of contextual diversity must consider semantic content. *Journal of Memory and Language*, *123*, 104313. https://doi.org/10.1016/j.jml.2021.104313

Keuleers, E. (2013). *vwr: Useful functions for visual word recognition research*. https://rdrr.io/cran/vwr/

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E.-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., & Scannell, K. (2020). *CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes—The National Corpus of Contemporary Welsh*. Cardiff University. http://doi.org/10.17035/d.2020.0119878310

Knight, D., Morris, S., Tovey-Walsh, B., Fitzpatrick, T., & Anthony, L. (2020). *Yr Amliadur: Frequency lists for contemporary Welsh* (Version 1.0.0). Cardiff University. http://doi.org/10.17035/d.2020.0120164107

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Brown University Press.

Kuipers, J.-R., & Thierry, G. (2010). Event-related brain potentials reveal the time-course of language change detection in early bilinguals. *NeuroImage*, *50*(4), 1633–1638. https://doi.org/10.1016/j.neuroimage.2010.01.076

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, Article 1171. https://doi.org/10.3389/fpsyg.2015.01171

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). *Assessment, testing and comparison of statistical models using R*. https://psyarxiv.com/vtq8f/

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics: System demonstrations* (pp. 55–60). https://aclanthology.org/P14-5010/

Martin, C. D., Dering, B., Thomas, E. M., & Thierry, G. (2009). Brain potentials reveal semantic priming in both the "active" and the "non-attended" language of early bilinguals. *NeuroImage*, *47*(1), 326–333. https://doi.org/10.1016/j.neuroimage.2009.04.025

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Mayr, R., & Davies, H. (2011). A cross-dialectal acoustic study of the monophthongs and diphthongs of Welsh. *Journal of the International Phonetic Association*, *41*(1), 1–25. https://doi.org/10.1017/S0025100310000290

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4). https://doi.org/10.1017/S014271640707035X

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52. https://doi.org/10.3758/BF03193811

Nieuwenhuis, R., Te Grotenhuis, M., & Pelzer, B. (2012). Influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. *R Journal*, *4*(2), 10.

Ooms, J. (2022). *Hunspell: High-performance stemmer, tokenizer, and spell checker*. https://hunspell.github.io

Prys, D., Prys, G., Jones, D. B., & Watkins, G. L. (2021). *Corpws CC0 Corpus* (21.10). Zenodo. https://doi.org/10.5281/zenodo.6376185

Prys, G., Prys, G., & Llewellyn, G. (2020). *Model Tagio Rhannau Ymadrodd Cymraeg* [Welsh language part of speech tagging model] (Version 20.10). https://github.com/techiaith/model-tagiwr-spacy-cy

Prys, G., & Watkins, G. (2022). Evaluation of three Welsh language POS taggers. In *LREC 2022 workshop language resources and evaluation conference 20-25 June 2022* (p. 30). https://www.research.ed.ac.uk/files/281144082/FransenEtal2022LREC2022Proceedings.pdf

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, *52*(6), 2372–2382. https://link.springer.com/article/10.3758/s13428-020-01389-1

van Heuven, W. J. B. (2020). *Pseudo* (Version 2.10) [Computer software]. https://waltervanheuven.net/pseudo/

van Heuven, W. J. B., 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*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vergara-Martínez, M., Comesaña, M., & Perea, M. (2017). The ERP signature of the contextual diversity effect in visual word recognition. *Cognitive, Affective, & Behavioral Neuroscience*, *17*, 461–474. https://doi.org/10.3758/s13415-016-0491-7

Watkins, G., Prys, G., & Jones, D. B. (2021). *techiaith/lecsicon-cymraeg-bangor: Lecsicon Cymraeg Prifysgol Bangor // Bangor University Welsh Language Lexicon (21.02)*. Zenodo. https://doi.org/10.5281/zenodo.5211667

Wu, Y. J., & Thierry, G. (2013). Fast modulation of executive function by language context in bilinguals. *The Journal of Neuroscience*, *33*, 13533–13537. https://doi.org/10.1523/JNEUROSCI.4760-12.2013

Yap, M. J., & Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 274–296. https://doi.org/10.1037/0278-7393.33.2.274

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971