The Auditory English Lexicon Project: A multi-talker, multi-region psycholinguistic database of 10,170 spoken words and nonwords

Winston D. Goh [1] · Melvin J. Yap [1] · Qian Wen Chee [1]

Abstract

The Auditory English Lexicon Project (AELP) is a multi-talker, multi-region psycholinguistic database of 10,170 spoken words and 10,170 spoken nonwords. Six tokens of each stimulus were recorded as 44.1-kHz, 16-bit, mono WAV files by native speakers of American, British, and Singapore English, with one from each gender. Intelligibility norms, as determined by average identification scores and confidence ratings from between 15 and 20 responses per token, were obtained from 561 participants. Auditory lexical decision accuracies and latencies, with between 25 and 36 responses per token, were obtained from 438 participants. The database also includes a variety of lexico-semantic variables and structural indices for the words and nonwords, as well as participants' individual difference measures such as age, gender, language background, and proficiency. Taken together, there are a total of 122,040 sound files and over 4 million behavioral data points in the AELP. We describe some of the characteristics of this database. This resource is freely available from a website (https://inetapps.nus.edu.sg/aelp/) hosted by the Department of Psychology at the National University of Singapore.

Keywords Speech database · Megastudy · Spoken word recognition · Auditory lexical decision · Psycholinguistics

The ease with which people are able to recognize printed and spoken words is one of the most impressive and important things humans do. Consequently, the processes underlying isolated word recognition and processing have been extensively studied (Balota, Yap, & Cortese, 2006; Dahan & Magnuson, 2006). Words are also one of the most commonly used set of stimuli in cognitive and experimental psychology (Balota et al., 2007). Researchers have accumulated a great deal of information regarding how the different statistical properties of words (e.g., frequency of occurrence, imageability, number of letters or phonemes) influence how quickly and accurately people can recognize words, and how they influence other cognitive processes, such as memory.

However, the overwhelming majority of experiments that have used word stimuli have focused on the processing of *printed* words. From a methodological point of view, the development and presentation of spoken, compared to printed, word stimuli is far more labor-intensive and complex. For example, each auditory token has to be recorded by one or more speakers, the sound file has to be edited to isolate the word, normalized, and tested for intelligibility before it can be used. In this light, it is perhaps unsurprising that empirical and theoretical developments in visual, compared to auditory, word recognition research have been relatively more rapid and extensive (see also Tucker, Brenner, Danielson, Kelley, Nenadić, & Sims, 2019). It is worth noting that Balota et al.'s (2007) English Lexicon Project's (ELP) behavioral and descriptive repository of visual word recognition data has contributed to these developments.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-020-01352-0) contains supplementary material, which is available to authorized users.

✉ Winston D. Goh
psygohw@nus.edu.sg

[1] Department of Psychology, National University of Singapore, Singapore 117570, Singapore

This paper describes the Auditory English Lexicon Project (AELP), which was conceived to address the above constraints by developing a very large and well-characterized set of auditory word and nonword tokens that have been rigorously normed for intelligibility. These tokens are freely available to the research community via a webpage (https://inetapps.nus.edu.sg/aelp/), and can be used for any experiment involving the presentation of spoken words and/or nonwords. In the following sections, we provide a brief overview of the theoretical importance of auditory word processing for understanding cognitive processes, existing spoken word databases, the megastudy approach and recent auditory megastudies, before turning to the AELP.

Auditory word processing

Listening and reading essentially have the same goal – retrieving the meaning of the stimulus, but effects do not always generalize across modalities, suggesting that there may be fundamental differences in the underlying mechanisms for lexical processing depending on the medium. For example, spoken word processing is consistently slowed down by dense phonological neighborhoods, but orthographic neighborhoods exert inconsistent effects in visual word recognition (Andrews, 1997). Semantic richness effects, the general finding that words with richer semantic representations facilitate processing (Pexman, 2012), have been shown to be smaller in auditory compared to visual word recognition (Goh, Yap, Lau, Ng, & Tan, 2016). These dissociations between visual and spoken word recognition point to the possibility that the recognition process in speech may focus more on resolving phonological similarities first (Goh, Suárez, Yap, & Tan, 2009; Luce & Pisoni, 1998), and so any advantages from semantically richer words are attenuated in the face of greater word-form competition.

Research has also shown that speech perception may be a talker-contingent process (Nygaard, Sommers, & Pisoni, 1994), and that indexical properties of spoken words – gender, accent, and other unique aspects of the talker's voice – are encoded and retained in memory (Goh, 2005; Goldinger, 1996b). Talker variability in the input enhances perceptual learning and word recognition in both adults (Logan, Lively, & Pisoni, 1991; Pisoni & Lively, 1995) and infants (Singh, 2008). These findings implicate the encoding of indexical information in long-term memory and provide support for an episodic mental lexicon (Goldinger, 1998).

In other cognitive domains, there is a well-known auditory advantage in the short-term memory (STM) literature, with several findings implicating the primacy of auditory codes in STM. For example, better memory for auditory compared to visually presented lists, especially in the primacy region, or the *modality effect* (Crowder, 1971; Penny, 1989); attenuation

of the recency effect if an irrelevant speech sound is played at the end of list presentation, or the *suffix effect* (Crowder & Morton, 1969); and fewer false memories for auditory versus visually presented lists of semantic associates (Olszewska, Reuter-Lorenz, Munier, & Bendler, 2015) but the reverse for phonological associates (Lim & Goh, 2019).

These selected examples highlight some of the important findings that differentiate studies using auditory versus visual stimuli, and studies using auditory tokens produced by multiple talkers. They point to the utility of having a large and easily accessible database of auditory tokens for experimental research.

Spoken word databases

As noted earlier, a significant bottleneck in auditory word recognition research has to do with the difficulty of developing auditory stimuli. The vast majority of existing speech databases comprise recordings of sentences, connected speech, and dialogue (e.g., TIMIT Acoustic-Phonetic Continuous Speech Corpus – Garofolo et al., 1993; The British National Corpus, 2007). These are generally not suitable for research using isolated spoken words. Some large isolated word databases tend to be tied to very specific contexts (e.g., 3000 names of Japanese railroad stations – Makino, Abe, & Kido, 1988). Hence, many researchers using auditory tokens prepare their own stimuli from scratch for most new studies.

In 2014, at the initial stages of the current project, there were no large spoken word databases readily available. Since then, three have been published and are summarized in Table 1.

Due to the time-consuming nature of creating large speech databases, it is unsurprising that all of these have a single talker recording a large number of tokens. Older collections tend to have many talkers saying a handful of words or sentences (e.g., the TIMIT corpus had recordings of 630 talkers speaking ten sentences each). The largest multi-talker database of English isolated spoken words thus far is the PB/MRT Word Multi-Talker Speech Database developed by the Speech Research Laboratory at Indiana University, comprising 450 words taken from phonetically balanced lists (IEEE, 1969) and the modified rhyme test (House, Williams, Hecker, & Kryter, 1965), and tokens of each word spoken by 20 American English talkers: ten male and ten female. To our knowledge, there is currently no multi-talker database of spoken nonword stimuli.

Auditory megastudies

Most insights on the nature of auditory word processing have come from factorial designs in which lexical and other

Table 1 List of large spoken word databases published after 2014

Source	Ernestus & Cutler, 2015	Ferrand et al., 2018	Tucker et al., 2019
Name	Biggest Auditory Lexical Decision Experiment Yet (BALDEY)	MEGALEX	Massive Auditory Lexical Decision (MALD)
Language	Dutch	French	English
No. of talkers	1	1	1
Gender	Female	Female (synthesized)	Male
No. of words	2780	17,876	26,793
No. of nonwords	2761	17,876	9592
No. of auditory lexical decision observations per token	20	101	4-6

variables of interest were manipulated while other properties were held constant. As more lexical properties were shown to affect spoken word recognition from standard factorial designs, it has become increasingly challenging to manipulate a single property while keeping others constant, and at the same time ensuring an adequate number of stimuli within each cell (Cutler, 1981). Turning continuous psycholinguistic variables into categorical variables (a prerequisite for factorial experimental designs) may also spuriously magnify or diminish the influence of variables and lessen the likelihood of detecting non-linear relationships (Balota, Yap, Hutchison, & Cortese, 2012).

This has precipitated a complementary research approach called the megastudy approach (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), where researchers allow the language to define the stimuli, rather than selecting stimuli based on a limited set of criteria. Specifically, participants are presented with a large set of words, and recognition performance for these words is measured. Statistical techniques are then used to estimate the unique influence of different targeted variables, while controlling for correlated variables. The megastudy approach has catalyzed the development of many large-scale databases across different languages (e.g., French, Spanish, Dutch, Malay, Chinese) and has also generated many productive lines of research in the visual modality, with the empirical findings critical for informing and constraining theories and models of reading and reading development (see Balota et al., 2012, for a review).

In the auditory domain, this megastudy approach has begun to be used, primarily with the lexical decision task (LDT), where participants have to discriminate between real words and pronounceable nonwords (e.g., *flirp*). The number of responses collected for each stimulus in auditory megastudies have ranged from 101 (MEGALEX) to 20 (BALDEY), and 4-6 (MALD). MEGALEX and BALDEY had participants listen to all stimuli in approximately 50 and ten sessions, respectively. MALD had a variable number of responses per token as listeners listened to 400 words and 400 nonwords in each session, and were allowed to participate up to a maximum of three such sessions (with different tokens) if they wanted to.

The Auditory English Lexicon Project

The AELP had three key objectives. First and foremost, we aimed to create a large database of spoken English words and nonwords that would be beneficial to all researchers requiring spoken word stimuli. To maximize its utility, the AELP was designed to be a multi-talker, multi-region database. It included six instances of each stimulus, as spoken by native speakers of American, British, and Singapore English, with one from each gender. The first two sets of talkers covered the world's two largest populations of native English speakers.

Second, it aimed to provide intelligibility norms for all tokens. Most spoken word recognition studies typically report that only tokens with some level of intelligibility (e.g., correctly identified by 80% of listeners) are used. None of the previous auditory megastudies have published any intelligibility scores for their tokens; only the accuracy rates in auditory lexical decision (ALD) were available. However, ALD accuracy data only indicate whether the token was perceived as a word, and not what word was heard. It is important to know what the correct identification rate for a spoken word is as it is possible that one word may be misheard for another. For example, *bag* may be misheard as *beg*, or vice-versa, but in both cases a participant would classify it as a word in an LDT, and would be scored as correct. However, the properties of the two words will differ (e.g., one is concrete and the other is abstract) and, depending on the intended target, the wrong properties may be used for subsequent experimental manipulation or analysis. Similarly, if a nonword is consistently heard as a word, it may not be a useful token to use in an experiment.

The third aim was to determine the robustness of word property effects for varieties of the same language across different talkers in ALD. All auditory megastudies have thus far been based on a single talker, and the extent to which effects can generalize across talkers, regions, and gender remains unknown.

The next section describes the stimuli selection considerations and summarizes the variables included in the database. The following two sections – Word recording and Word identification – describe the methodologies and results specific to the first two aims. These two phases of the project served to iteratively record the stimuli and test their intelligibility in order to finalize the list of tokens to be included in the database. The section on Word recognition will describe the behavioral data that are available and addresses the third aim. All protocols were approved by the National University of Singapore (NUS) Institutional Review Board. The final sections will describe some analyses and examples of how the data may be used, and a brief overview of the features of the website.

Stimulus selection

Words

The goal of the AELP was to maximize the utility of the word stimuli for as many researchers as possible. Two general principles were adopted to achieve this: one, the words must be familiar to most people, and two, the words should have values or ratings on as many psycholinguistic variables and word properties that are currently available in the literature.

Although printed word frequency and subjective familiarity ratings covary, the latter has been shown to be a more reliable predictor of lexical decision performance (e.g., Gernsbacher, 1984; Kreuz, 1987). To the best of our knowledge, the largest database of subjective familiarity (FAM) ratings thus far is Nusbaum, Pisoni, and Davis' (1984) Hoosier Mental Lexicon (HML), which collected ratings for 19,750 words from the intersection of Merriam-Webster's Pocket Dictionary and Webster's Seventh Collegiate Dictionary. Each word's FAM score was derived from averaging 12 observations, based on a seven-point scale, where a rating of 1 indicated an unknown word, a 4 indicated that the rater knew the stimulus was a word, but did not know the meaning, and a 7 indicated that the word was recognized with at least one meaning known. The other points represented intermediate levels of familiarity (the full rating scale is listed in Appendix 1).

The aim was to have around 10,000 words in the AELP. Words with Hoosier FAM scores of at least 6.2 were selected to form the initial list of 10,446 words. We cross-checked this with words that had NUSFAM[1] scores of at least 6, and removed function words (e.g., *am*), and people's names. Several research assistants went through the list and highlighted unusual words (e.g., *choler*, *clew*), which were eventually dropped. All words not already included from the McRae, Cree, Seidenberg, and McNorgan (2005) number of features (NoF) norms were included in order to optimize the use of this semantic property. The final AELP database comprised 10,170 words[2].

Nonwords

Unique nonwords were created and yoked to each word in the database so that they resembled the target word as closely as possible. Although the WUGGY pseudoword generator (Keuleers & Brysbaert, 2010) was developed to create orthographically rather than phonologically plausible words, it was used as an initial guide to list plausible candidates for selection or modification. We followed BALDEY's approach in ensuring that, as far as possible, the structure of the tokens became nonwords only towards the end of the phoneme sequence so that nonword detection cannot be strategically determined. Specifically, for words with three or more syllables, at least the first syllable was retained in its entirety (e.g., for the word *orchestra* /ˈɔrˌkɛstrə/, a WUGGY suggested nonword was *orshistre*, which was then modified to *orchistro* and transcribed as /ˈɔrˌkɪstroʊ/, thus retaining the first syllable of the

---

[1] At the time of stimulus selection, our laboratory had also collected FAM ratings for 6117 words with scores derived from 25–133 observations per word, for various research and student projects since 2002, using the same scale as the HML. This set of 6117 was used to cross-check the Hoosier FAM words. At the end of this project, NUSFAM grew to 12,163 words, including all words in the AELP.

[2] The final set included a few changes based on outcomes from the Word identification phases.

target word and retaining two-thirds of the phonemes in the other syllables to derive the nonword).

It was more challenging to follow this principle for monosyllabic and disyllabic words. For both sets, we tried to retain the first phoneme of each syllable, in order to ensure that about one- to two-thirds of the target word's phonemes were retained. However, it was not always possible if all candidate nonwords were already yoked to other words (e.g., the nonword *zoong* /zuŋ/ had to be yoked to the word *earn* /ə：n/).

Descriptive characteristics

Tables 2 and 3 summarize the structural and lexico-semantic variables, respectively, that are included in the database. The descriptive statistics are included in Appendix 2.

Structural indices such as neighborhood density were computed separately for American and British English. We used the Oxford Dictionaries (2019) Application Programming Interface (API)[3] to obtain International Phonetic Alphabet (IPA)[4] transcriptions of 26,604 of the 40,481 ELP (Balota et al., 2007) words that had phonological transcriptions in *both* American and British English[5]. This base dataset is close in number to the 20,000 words in the HML, which has also been used to compute phonological indices in many previous studies (see Vitevitch, 2008) and so would allow a similar base for comparison of the indices with previous work. The structural properties of the 10,170 AELP words and nonwords were computed with reference to the 26,604 base dataset, which is included in the supplemental material.

Table 4 summarizes the correlations between some of the structural phonological properties for words and nonwords, and indicates that the nonwords resemble the yoked words closely and share similar properties, as intended.

From Table 19 in Appendix 2, most of the words in the AELP have the relevant measures from other large databases, ranging from 75% (7612/10170) for the Warriner et al. (2013) affective norms to 98% (10012/10170) for the Brysbaert et al. (2019) prevalence norms. We then checked the extent to which other well characterized lists of words that have been used extensively in the field are found in the AELP database

(Table 5). From both Tables 5 and 19, it can be seen that the AELP database should have enough words to be an optimal resource for selecting auditory experimental stimuli in psycholinguistic and cognitive studies.

Word Recording

Talker Selection

Potential talkers were recruited from the NUS and wider expatriate communities. To minimize regional dialect differences for the American and British talkers, we considered only those who grew up (for the first 18 years of their lives) in the mid-western states of the United States (for American talkers) or in the Home Counties (for British talkers). Experience with and the ability to read the International Phonetic Alphabet (IPA) were also emphasized.

Fifteen participants were invited to an individual 1-h trial recording session using the apparatus and procedures described in the next section, and were each reimbursed S$10. A list of 80 words from Goh et al. (2016), for which intelligibility norms were available from that study, plus the corresponding yoked nonwords from the AELP, was used in the trial session. For each of the seven participants with the clearest recordings and who could potentially serve as talkers, their 80 word and 80 nonword tokens were presented to seven different groups of 15–20 undergraduate NUS students each for intelligibility testing (details are described in the Word identification section). All seven potential talkers exceeded the average correct identification (ID) rate of .76[6] for the same 80 words from Goh et al. (2016), but we dropped one female Singapore English speaker to form the final six AELP talkers, whose characteristics are summarized in Table 6.

Recording and editing procedures

All recordings were done in a bespoke sound isolation booth with a Field Sound Transmission Class (FSTC) 56 rating. Speech signals were captured with an Audio-Technica ATM75 cardioid condenser head-mounted microphone connected to a Pentax Computerized Speech Lab (CSL) Model 4500 voice recorder, and saved as 16-bit mono, 44,100-Hz .wav sound files.

---

[3] As Oxford Dictionaries declined permission, we are unable to include the words' phonological transcriptions in the database. Nonetheless, interested researchers are able to request free researcher API access from Oxford Dictionaries, and we have included the annotated Python code used to retrieve the transcriptions in the supplemental material.

[4] A description of the IPA symbols used by Oxford Dictionaries is available at https://public.oed.com/how-to-use-the-oed/key-to-pronunciation/ (retrieved 16 October 2019).

[5] A dictionary for Singapore English is currently unavailable at Oxford Dictionaries. Users who want estimates of Singapore English should use the British metrics. The pronunciation of Standard Singapore English, which is the variety used by our Singaporean talkers, is closer to British than American English (e.g., *ask* is pronounced /ɑːsk/ and not /æsk/; *car* is pronounced /kɑː/ and not /kɑr/), and spelling follows British conventions (see Deterding, 2007; Leimbgruber, 2011; Tay & Gupta, 1983).

[6] This rate is relatively low as we deliberately chose several words in the list of 80 that had low ID scores, e.g., *colander*, ID = .23, in order to see if potential talkers could be clearer than the one used in the earlier study. The average ID rate for all words for the talker in Goh et al. (2016) is much higher at .94.

Table 2 Structural variables included in the database

Property	Description		
	Orthographic	Phonological	Phonographic
Word length	Number of letters	n/a	n/a
Number of phonemes	n/a	Phoneme count	n/a
Number of syllables	n/a	Syllable count	n/a
Neighborhood density	Number of words formed by a single letter change	Number of words formed by a single phoneme change	Number of words formed by a simultaneous single letter and single phoneme change (Peereman & Content, 1997)
Neighborhood frequency	Average word frequency (using LgSUBTLWF) of the target word's neighbors		
Identity of neighbors	List of orthographic neighbors	List of phonological neighbors	List of phonographic neighbors
Substitutions only (computed for density, neighborhood frequency, and identity of neighbors)	Replacing one letter of the target word only, e.g., *hat*, *cut*, and *cap* are neighbors of *CAT*	Replacing one phoneme of the target word only, e.g., /hæt/, /kʌt/ and /kæp/ are neighbors of *CAT*	Replacing one letter and one phoneme of the target word only, e.g., *eat*, *oaf*, and *out* are neighbors of *OAT*
Substitutions, additions, and deletions (computed for density, neighborhood frequency, and identity of neighbors)	Includes adding or deleting a single letter, e.g., *scat* and *at* are also neighbors of *CAT*	Includes adding or deleting a single phoneme, e.g., /skæt/ and /æt/ are also neighbors of *CAT*	Includes adding or deleting a single letter and single phoneme, e.g., *boat* and *at* are also neighbors of *OAT*
Spread	Number of letter positions that can be substituted to form a neighbor (Johnson & Pugh, 1994)	Number of phoneme positions that can be substituted to form a neighbor (Vitevitch, 2007)	n/a
Uniqueness point	Point where the word diverges from all other words in the lexicon, measured in terms of number of letters starting from the initial letter (Marslen-Wilson & Welsh, 1978)	Point where the word diverges from all other words in the lexicon, measured in terms of number of phonemes starting from the initial phoneme (Luce, 1986)	n/a
Levenshtein distance (OLD/PLD-20)	Average edit distance of the target word's 20 closest neighbors (Yarkoni, Balota, & Yap, 2008)		n/a
C coefficient	Degree to which neighbors, based on single letter substitution, addition, or deletion, of a target word are also neighbors of each other (Watts & Strogatz, 1998)	Degree to which neighbors, based on single phoneme substitution, addition, or deletion, of a target word are also neighbors of each other (Watts & Strogatz, 1998)	Degree to which neighbors, based on simultaneous single letter and single phoneme substitution, addition, or deletion, of a target word are also neighbors of each other
Summed bigram/biphone frequency	The position-specific summed log frequency of bigrams	The position-specific summed log frequency of biphones (Vitevitch & Luce, 2004)	n/a

Phonological structural metrics were also computed for nonwords

Table 3 Lexico-semantic variables included in the database

Property	Description
Number of morphemes	Number of meaningful units within a word
Part of speech (POS): adjective, determiner, noun, pronoun, adverb, interjection, number, verb, conjunction, name, preposition	Based on the SUBTLEX databases
Word frequency	Log-transformed subtitle word frequency estimates (Brysbaert & New, 2009)
Contextual diversity	Log subtitle contextual diversity estimates (Brysbaert & New, 2009)
Familiarity (FAM)	Average subjective familiarity ratings
NUS FAM	From National University of Singapore students
Hoosier FAM	From Indiana University students (Nusbaum et al., 1984)
Word prevalence	Estimates of the number of people who know the word (Brysbaert, Mandera, McCormick, & Keuleers, 2019)
Prevalence	
Percentage known	
Concreteness rating	Measure of word concreteness (Brysbaert, Wariner, & Kuperman, 2014)
Age of acquisition (AoA)	Index of age when the word is acquired (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012)
Semantic neighborhood density (SND)	Measures of semantic neighborhood density (Shaoul & Westbury, 2010)
Neighbor count (Ncount)	Number of neighbors within a pre-specified neighborhood threshold
Average radius of co-occurrence (ARC)	Mean distance between target word and its neighbors within a pre-specified neighborhood threshold
Affective ratings	Average subjective ratings of affect (Warriner, Kuperman, & Brysbaert, 2013)
Valence	How pleasant the word is (Osgood, Suci, & Tannenbaum, 1957)
Arousal	Degree to which the word evokes arousal (Osgood et al., 1957)
Dominance	Extent to which the word denotes weakness/submissiveness or strength/dominance (Osgood et al., 1957)

The words and nonwords, excluding the 80 words and 80 nonwords already recorded in the trial recording sessions, were divided into 40 lists of 252 token-sets each[7]. A token-set comprised a word and its corresponding nonword. The lists were equated on three lexical properties: the average log-transformed subtitle word frequency (LgSUBTLWF) from Brysbaert and New (2009), the average number of phonemes (N_Phon), and the average phonological Levenshtein distance (PLD20), which indexes the mean number of phoneme deletions, insertions, or substitutions required to transform a word into its 20 closest Levenshtein neighbours (Yarkoni et al., 2008), all p values > .19. This was done to minimize between-list differences in potential pronunciation difficulty arising from frequency of use, word length, and word-form similarity, respectively.

For each token-set, the talker produced two utterances of the word with a pause between each utterance, followed by a similar two-utterance sequence for the nonword, aided by its IPA transcription. Each talker completed about 85 token-sets before the research assistant (RA) saved the sequence of utterances in a raw, uncut .wav file. Any mispronunciations or dysfluencies noted by the RAs, who were all trained in IPA phonetic notations, were re-recorded. One list could be recorded in an hour, with each talker recording for between two and

---

[7] The original intention was to have 10,160 words in the database. Some words were replaced eventually because of poor intelligibility and new words added for recording in the subsequent re-recording sessions.

Table 4 Correlations between words and nonwords' structural phonological properties

Property	American English	British English
Number of phonemes	.97	.97
Number of syllables	.97	.99
Neighborhood density (substitution, addition, and deletion)	.73	.75
PLD20	.88	.90

four hours per session, with multiple breaks as needed. Talkers were reimbursed S\$60 for every 1000 token-sets, and a completion bonus of S\$400 was provided when all tokens were recorded. All talkers completed the full set of recordings in 11 to 16 sessions. For re-recordings after the ID phases, the talkers were reimbursed at the same rate as above, but without the completion bonus.

Adobe Audition was used to edit the raw recordings to be saved as individual .wav files for each of the utterances. All tokens were then digitally levelled to approximately 70 dB to ensure that all tokens had the same total root-mean-square amplitude using the match loudness function in Audition. RAs then indicated which of the two instances of each token was better, based on enunciation, noise, and realization of vowels. The better token was subsequently used in the ID phases.

Word identification

This phase of the project served to determine the intelligibility of the selected tokens and whether re-recordings were required for some items. Tokens that did not achieve an ID rate of at least .75 were re-recorded (or changed). The new tokens were then subjected to another round of intelligibility testing.

Materials

The first round of word ID (WordID1) comprised the 80 token sets (words and nonwords) from the trial recording sessions. The second round (WordID2) included the remaining 10,080 token-sets. This was divided into 30 lists[8] of 336 token-sets and grouped into six groups of five lists each, with all lists and groups equated on the three lexical properties for the words as in the recording sessions, all Fs < 1. The third round (WordID3) comprised re-recordings of tokens that failed to achieve the ID criterion in both WordID1 and WordID2,

which varied between 10.24% and 11.66% of tokens across talkers, and were divided into three lists of varying numbers of tokens. This round also included testing of some new non-words and words for possible replacement of tokens that had consistently poor ID rates.

Participants

The goal was to have each token tested by between 15 and 20 different participants. A total of 561 participants from the NUS community took part in one or more rounds of the identification phase. It was ensured that no participant heard each token more than once within or across talkers. In their first session, participants completed a language background questionnaire (LBQ) before the ID task, and the 40-item vocabulary subscale of the Shipley Institute of Living Scale (Shipley, 1940) at the end of the session.

Participants were reimbursed S\$5 for every half-hour. WordID1 comprised a single half-hour session; WordID2 was run as cycles of five 1-h sessions, with participants having an option of participating in more than one cycle; and WordID3 had three 1-h sessions. For WordID2, adapting from MEGALEX's completion incentive, participants were given a S\$25 bonus for every five sessions completed; and adapting from MEGALEX's continuation criterion, participants were also told that they would be dropped from further sessions if their accuracy rate dropped below 80% for two consecutive sessions.

Table 7 summarizes the profile of the participants. All participants indicated English as their first language, reported no speech or hearing disorder, and had lived in Singapore for more than half their lives.

Procedure

Participants were tested in groups of 12 or fewer on individual PCs running E-prime (Schneider, Eschman, & Zuccolotto, 2002). On each trial, an auditory token was binaurally presented via beyerdynamic DT150 headphones at approximately 70 dB SPL. Participants were instructed to make a judgment on whether what they heard was a word or nonword. For words, participants had to spell out the word using the keyboard; for nonwords, they typed an "x". They pressed the ENTER key to submit the trial response, after which they were asked to rate the confidence of their response on a nine-point Likert scale, with higher numbers representing greater confidence. Pressing a number initiated the next trial after a 100-ms blank screen.

For the multi-session cycles in Word ID2, each 1-h session was done approximately 1 week apart. One list was presented in each session and word and nonword tokens were randomly interspersed for each participant. A balanced latin-square was used to rotate the order of lists within each cycle across the

---

[8] Talker F1's WordID2 was based on the 40 lists from the main recording session, but most participants were able to complete a list of 504 tokens (252 words and corresponding nonwords) in 30–40 minutes in one session, and so we decided to increase the number of tokens in one session for the remaining talkers.

Table 5 Proportion of AELP words with measures from smaller word property databases

Property	Source	No. of words in source	No. of words in AELP	Percent of source in AELP (%)
Imageability	Cortese & Fugett (2004)	6277	4776	76.09
Imageability	Schock, Cortese, & Khanna (2012)	3000	2272	75.73
Number of features (NoF)	McRae et al. (2005)	541	491	90.76
Phonological false memory associates	Sommers & Lewis (1999)	744	552	74.91
Semantic false memory associates	Roediger, Watson, McDermott, & Gallo (2001)	825	735	89.09

sessions. A similar procedure was used for the rounds in WordID3. Participants were given a break every 75–90 trials in each session of WordID2 and WordID3. There was one break after 80 trials for WordID1. Participants were debriefed at the end of their last session.

Scoring

For words, all incorrect responses that were automatically flagged by E-prime based on spelling accuracy were checked to see if they could be considered correct. For example, obvious spelling errors (*occurrence* spelt as *occurence*), typographical errors involving adjacent letters on the keyboard (*violin* spelt as *violim*), British/American spelling variants (*colour* spelt as *color*), and homophone responses (*sail* spelt as *sale*), were all re-scored as correct for the purposes of estimating ID rates[9]. For nonwords, all incorrect responses were checked to see if any nonword was consistently perceived to be a word (e.g., /tʃəp/ was heard as *chop*). These were either re-recorded or a new nonword was recorded and tested again.

At the end of WordID3, all tokens were finalized regardless of their ID rate. For tokens that went through two rounds of testing (WordID1/WordID2 and WordID3), the token with the better ID was kept in the database. Table 8 summarizes the average ID and confidence rates for the talkers, and Table 9 depicts the correlations between the ID rates for each talker. The relatively high intercorrelations attest to the high reliability of the intelligibility measure (Keuleers, Lacey, Rastle, & Brysbaert, 2012) and reflect the amount of item-level variance that can be explained in these datasets (Courrieu & Rey, 2011).

Figure 1 depicts the proportion of tokens across different ID rates among the six talkers. The results show that the vast majority of tokens (over 86%) have an ID score of at least .8, indicating that the AELP database comprises highly intelligible tokens that can be used in experiments.

---

[9] Individual homophone ID rates are described in Appendix 4.

Word recognition

This phase of the project collected latency and accuracy data for auditory lexical decision for all tokens and talkers.

Materials

The AELP words were divided into 15 lists of 678 token-sets and grouped into three groups of five lists each, with all lists and groups equated on the three lexical properties for the words as in the recording sessions, all $Fs < 1$.

Participants

The goal was to have each token tested by between 25 and 36 different participants. Four hundred and thirty-eight participants from the NUS community took part and it was ensured that no participant heard each token more than once within or across talkers. As in the ID phase, in their first session, participants completed a language background questionnaire (LBQ) before the LDT. This time, we replaced the Shipley (1940) subscale with a newer 60-item subset of a vocabulary test (Levy, Bullinaria & McCormick, 2017) and a 60-item spelling test (Burt & Tate, 2002) at the end of the first session.

Table 6 Talker characteristics

Talker code	English variety	Age at time of recording	Region grew up in for first 18 years
Female			
F1	Singapore	22	Singapore
F2	American	21	Minnesota
F3	British	38	Surrey
Male			
M1	Singapore	27	Singapore
M2	American	21	Wisconsin
M3	British	37	Essex

Table 7 WordID participant profile

	Female (n = 391)		Male (n = 170)	
	M	SD	M	SD
Age	20.60	1.34	22.42	1.41
Vocabulary score	30.26	3.38	32.07	3.33

Most (44.92%) participants completed five sessions of WordID2. 21.03% completed ten sessions, 5.35% completed 15 sessions, 0.89% completed 20 sessions, and 0.89% completed 25 sessions; 12.48% took part only in WordID1 and 14.44% took part only in WordID3

Similar to the ID phase, participants were reimbursed S$5 for every half-hour and a S$25 bonus for every five sessions completed. The same continuation criterion was adopted; participants were told that they would be dropped from further sessions if their accuracy rate dropped below 80% for two consecutive sessions.

Table 10 summarizes the profile of the participants. All participants indicated English as their first language, reported no speech or hearing disorder, and had lived in Singapore for more than half their lives. Other language details can be found in Appendix 4.

Procedure

Participants were tested in groups of 12 or fewer on individual PCs running E-prime with the Chronos response box (Schneider, Eschman, & Zuccolotto, 2012) that had the leftmost and rightmost buttons labelled "nonword" and "word" respectively. On each trial, an auditory token was binaurally played via beyerdynamic DT150 headphones at approximately 70 dB SPL. Participants were asked to determine, as quickly and as accurately as possible, whether the token was a word or a nonword. Latency was measured from stimulus onset till the button press. An inter-stimulus-interval (ISI) of 200 ms elapsed before the next token was played. Participants were given a short break after every 113 trials.

Table 8 Intelligibility rates across talkers and stimuli

Talker	Word				Nonword			
	ID		Confidence		ID		Confidence	
	M	SD	M	SD	M	SD	M	SD
F1	.93	.12	8.25	0.54	.92	.10	6.26	0.66
F2	.92	.13	8.18	0.58	.93	.12	6.41	0.63
F3	.92	.14	8.25	0.62	.92	.10	6.33	0.63
M1	.92	.14	8.13	0.64	.94	.10	6.34	0.55
M2	.91	.14	8.34	0.56	.91	.12	6.42	0.66
M3	.92	.13	8.26	0.61	.93	.10	6.36	0.60

As in WordID, each 1-h session was done approximately 1 week apart. One list was presented in each session and word and nonword trials were randomly interspersed for each participant. A balanced latin-square was used to rotate the order of lists within each group across participants and sessions. Participants were debriefed at the end of their last session.

Behavioral measures

Response times (RT) were cleaned as follows. First, all inaccurate responses were removed (10.64% for words, 8.40% nonwords). Next, RTs less than 200 ms and greater than 3000 ms were removed (0.42% words, 0.85% nonwords). RTs that were greater or less than 2.5 SDs from each individual participant's overall mean for that session were then removed (1.74% words, 3.20% nonwords). We also computed a dependent measure that subtracted token duration from RT, which has been used in some previous studies (e.g., Luce & Pisoni, 1998; Taft & Hambly, 1986; see also Goldinger, 1996a).

All valid responses were then averaged for each item to obtain item-level measures for each talker. We also computed dialect-level estimates, which comprised trials averaged across both male and female talkers of the same dialect, using the same logic as described earlier. The behavioral measures available in the database are listed in Table 11. The descriptive statistics are summarized along with the durations of the finalized tokens in Appendix 3.

The correlations between talkers and dialects for zRT are summarized in Tables 12 and 13. The standardized item score is a more reliable measure of LDT performance as different sets of participants contributed RTs for different words, and the standardized score minimizes the influence of a participant's processing speed and variability (see Faust, Balota, Spieler, & Ferraro, 1999). The full correlation matrix between behavioral measures and the structural and lexico-semantic properties can be found in the supplemental materials.

Sample analyses and uses of the data

Intelligibility data

Speech intelligibility has traditionally been studied using the perceptual identification task, where tokens are identified at different signal-to-noise ratios, using noise as a mask (see Pisoni, 1996), and seldom in the clear. Word frequency and familiarity facilitates, whereas similarity in lexical neighborhoods inhibits, correct identification in noise (e.g., Rosenzweig & Postman, 1957; Treisman, 1978). We explored the influence of structural and lexical variables on our ID data, which was presented without noise, by performing a multiple regression for each talker with ID scores as the criterion and two measures of

Table 9 Correlations between talker intelligibility rates for words (*upper triangle*) and nonwords (*lower triangle*)

Talker	F1	F2	F3	M1	M2	M3
F1		.60	.66	.68	.62	.66
F2	.44		.68	.67	.72	.66
F3	.42	.51		.74	.70	.77
M1	.48	.49	.46		.69	.73
M2	.46	.63	.49	.49		.70
M3	.47	.48	.54	.51	.49	

familiarity (NUSFAM and prevalence), two measures of phonological similarity (neighborhood density and PLD20), and a measure of word frequency (lgSUBTLCD) as predictors. The outcome is summarized in Table 14.

The results are highly consistent across talkers. More familiar and more frequently occurring words have higher intelligibility, while more phonologically confusable words (higher density and closer PLD20 distances) are associated with poorer intelligibility. These patterns indicate that although only high FAM words were included in the database, there is still enough range in the database to facilitate the exploration of lexical and other influences on behavioral outcomes.

Auditory Lexical Decision Data

Item-level regressions To determine whether the database could replicate some of the classic findings in the auditory lexical decision literature, such as word frequency facilitation (e.g., Taft & Hambly, 1986) and phonological neighborhood
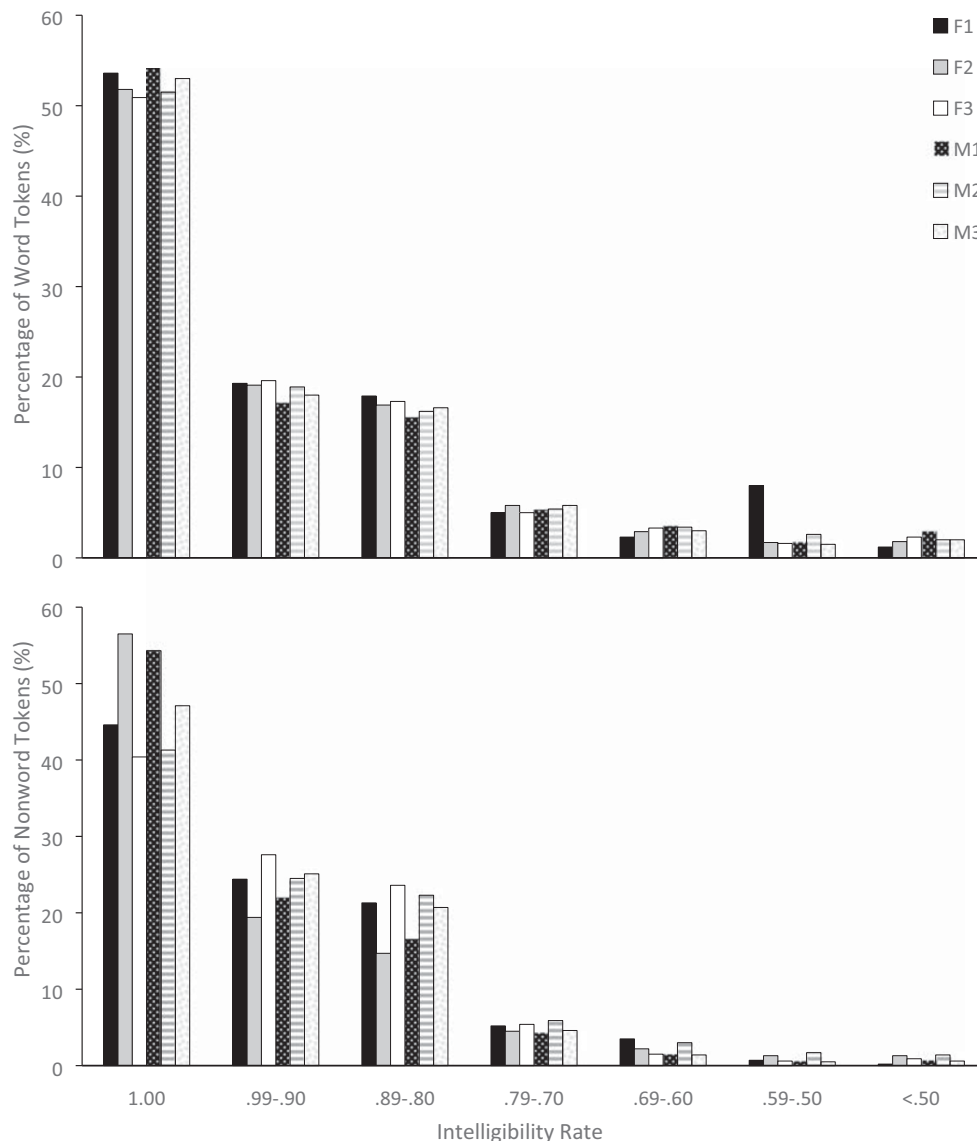

Fig. 1 Word (*top panel*) and nonword (*bottom panel*) ID rate percentages across talkers

Table 10 Word recognition participant profile

	Female (*n* = 298)		Male (*n* = 140)	
	M	*SD*	*M*	*SD*
Age	20.50	1.26	22.34	1.36
Spelling score	54.91	3.11	54.39	3.44
Vocabulary score	45.52	5.22	48.32	4.81

Most (79.00%) participants completed five sessions of word recognition; 20.32% completed ten sessions, and 0.68% completed 15 sessions

density competition (e.g., Luce & Pisoni, 1998), item-level multiple regression analyses were performed for the zRT word estimates from each talker. As there were very high correlations between number of phonemes, number of syllables, and PLD20 (|*r*|s between .82 and .92), principal components analysis (PCA) was used to reduce these to a single component, separately for the American and British English values; varimax rotation with Kaiser normalization was used. The component accounted for 92% of the variance, with higher values indicating greater phonological distinctiveness. Table 15 depicts the component loadings.

Words with estimates based on fewer than ten observations (indicating a low accuracy rate), were then dropped. Table 16 summarizes the regression results. Positive relationships were found between zRT and number of morphemes, token duration, and neighborhood density. Words with longer tokens, had more morphemes, or resided in dense neighborhoods (more confusable) were associated with slower latencies. Negative relationships were found between zRT and familiarity, prevalence, frequency, and the principal component. Words that were more familiar, prevalent, encountered more often, and more phonologically distinct (less confusable) were associated with faster RTs.

Linear mixed effects modeling We also analyzed the data using a linear mixed effects (LME) model for the same variables, using R (R Core Team, 2019). zRTs were fitted using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015);

Table 11 ALDT talker- and dialect-level behavioral estimates available for each item in the database

Measure	Description
Accuracy	Response accuracy
Reaction time	Raw response latency
z-score reaction time	z-score response latency, by subject and by session
z-score (reaction time minus duration)	z-score (response latency minus token duration), by subject and by session

Talker-level estimates for intelligibility and confidence ratings from the WordID phases are also available as behavioral measures, as described in the previous section

Table 12 Accuracy and response latency correlations between talkers for words (*upper triangle*) and nonwords (*lower triangle*)

Talker	F1	F2	F3	M1	M2	M3
Accuracy						
F1		.75	.76	.76	.75	.73
F2	.54		.77	.74	.80	.75
F3	.57	.59		.76	.78	.80
M1	.57	.54	.56		.75	.74
M2	.53	.66	.59	.54		.77
M3	.58	.53	.62	.60	.56	
zRT						
F1		.72	.72	.69	.72	.67
F2	.59		.76	.72	.74	.68
F3	.62	.67		.72	.72	.73
M1	.59	.60	.64		.68	.64
M2	.59	.61	.60	.58		.70
M3	.55	.57	.64	.53	.60	

p values for fixed effects were obtained using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2016). The influence of the variables was treated as fixed effects. Random intercepts for participants and items, and random slopes for frequency (lgSUBTLCD), NUSFAM, phonological neighborhood density, and PCA were included in the model.

As can be seen in Table 17, the pattern of results converges with those obtained in the item-level regression analyses. Inhibitory effects on zRT were observed for number of morphemes, token duration, and neighborhood density, while facilitatory effects on zRT were observed for familiarity, prevalence, frequency, and the principal component.

Taken together, these patterns replicate the more robust findings in auditory lexical decision, such as facilitation for more frequent words and competition between similar sounding words. They are also consistent with findings from visual lexical decision for word prevalence (Brysbaert et al., 2019), which has not been previously explored in the auditory domain.

Table 13 Accuracy and response latency correlations between dialects for words (*upper triangle*) and nonwords (*lower triangle*)

Dialect	Singapore	American	British
Accuracy			
Singapore		.84	.84
American	.64		.85
British	.71	.67	
zRT			
Singapore		.82	.80
American	.73		.82
British	.72	.73	

Table 14 Intelligibility standardized regression coefficients for item-level regression analysis across talkers

Variable	F1	F2	F3	M1	M2	M3
NUSFAM	.43***	.43***	.51***	.49***	.44***	.46***
Prevalence	.15***	.16***	.18***	.18***	.18***	.20***
Phonological neighborhood density (substitution, addition, and deletion)	-.15***	-.12***	-.10***	-.13***	-.10***	-.10***
PLD20	.11***	.12***	.10***	.11***	.13***	.12***
lgSUBTLCD	.06***	.05***	.05***	.03**	.08***	.05***
Adjusted R^2	**.33***	**.34***	**.43***	**.41***	**.37***	**.40***

** $p < .01$, *** $p < .001$

Virtual factorial experiments

We have already described some regression and lme analyses based on the megastudy approach in the earlier sections, for both word identification and auditory lexical decision. The database can also be used to do auditory lexical decision virtual experiments for smaller word lists, such as those found in factorial experiments. One important feature of the AELP multi-talker database is that different participants heard different tokens across the six talkers. Hence, we can essentially do a total of three to six virtual replications, depending on whether dialect-level or talker-level estimates are used. An obvious advantage of this is that multiple replications would provide more confidence on the robustness and generalizability of any observed effects.

In the following sections, we describe three such virtual experiments on a lexical property (word frequency), a structural property (phonological onset-density), and a semantic property (number of features) that have been reported in the literature.

Word frequency Taft and Hambly (1986) reported that high frequency words were responded to more accurately and faster than low frequency words in their Experiment 4. We performed a 2×6 mixed-design analysis of variance (ANOVA) using zRT as the dependent variable, with Frequency (high, low) as the between-items factor and Talker (F1 to M3) as the within-items factor, on 38 of their 48 words found in the AELP database. The frequency effect was replicated and depicted in Fig. 2. High-frequency words were associated

with more accurate, $F(1, 36) = 11.79$, $MSe = .04$, $p < .01$, and faster responses, $F(1, 36) = 14.42$, $MSe = .36$, $p < .01$, than low-frequency words.

Onset density Vitevitch (2002) reported that words with few neighbors sharing the initial phoneme (sparse onset density) were responded to more quickly than words with many neighbors sharing the initial phoneme (dense onset density) in his Experiment 2, with no differences in accuracy. We performed a 2×6 mixed-design ANOVA with Onset Density (sparse, dense) as the between-items factor and Talker as the within-items factor, on 84 of his 90 words found in the AELP database. The onset density effect was replicated and depicted in Fig. 3. Sparse-onset words were associated with faster responses, $F(1, 82) = 7.47$, $MSe = .65$, $p < .01$, than dense-onset words, with no differences in accuracy, $F < 1$.

Number of features Sajin and Connine (2014) reported that words with high number of features (NoF) were responded to more quickly and accurately than words with low NoF in their Experiment 1. We performed a 2×6 mixed-design ANOVA with NoF (low, high) as the between-items factor and Talker as the within-items factor, on all 115 words found in the AELP database. The NoF effect was replicated and depicted in Fig. 4. High NoF words were associated with more accurate, $F(1, 113) = 8.67$, $MSe = .05$, $p < .01$, and faster responses, $F(1, 113) = 8.23$, $MSe = .69$, $p < .01$, than low NoF words.

In the future, when additional word properties are invented or discovered, users could use the RT data in the database to conduct more of such virtual experiments in tandem with running actual factorial designs, in order to obtain converging evidence and replications for hypothesized effects.

The AELP website

Figure 5 shows the homepage of the database's website. It is separated into three main sections. The Generate section allows users to generate a list of words from the database with user-specified lexical properties and behavioral data. The

Table 15 Principal component loadings

Predictors	Loading	
	American English	British English
Number of phonemes	.98	.98
Number of syllables	.95	.96
PLD20	.95	.95

Table 16   zRT standardized regression coefficients for item-level regression analyses across talkers

Variable	F1	F2	F3	M1	M2	M3
Number of morphemes	.07***	.07***	.06***	.07***	.07***	.08***
NUSFAM	− .23***	−.21***	−.16***	−.16***	−.20***	−.18***
Prevalence	− .08***	−.11***	−.09***	−.08***	−.09***	−.07***
Phonological neighborhood density (substitution, addition, and deletion)	.21***	.18***	.19***	.17***	.19***	.21***
lgSUBTLCD	−.17***	−.13***	−.14***	−.11***	−.17***	−.14***
PCA	−.14***	−.04***	−.07***	−.04**	−.01	−.07***
Duration	.68***	.69***	.80***	.74***	.59***	.69***
Adjusted R^2	**.55***	**.62***	**.68***	**.62***	**.53***	**.57***

** $p < .01$, *** $p < .001$

Submit section allows users to submit a list of words to obtain user-specified lexical properties and behavioral data from the database. Users can also download the relevant sound files from these two sections. The final Downloads section allows users to download all sound files, trial-level lexical decision data, and a few other useful lists.

Generate

Figure 6 illustrates part of the Generate section page with a user-specified example. Here, the user wants a list of five- and six-syllable words and nonwords for British English from the database, and to include IPA pronunciations for the latter. The user would check the relevant properties, and can use the slider to limit the range for the number of syllables property in the phonological metrics section. The user then clicks the download button at the bottom of the page and the relevant properties will be retrieved as a .csv file named aelp_data.csv, stored in a .zip file called aelp.zip. Figure 7 shows the csv file as opened in Microsoft Excel for the example described.

Submit

Figures 8 and 9 illustrate part of the Submit section page with a user-specified example. Here, the user inputs five words for

American English and wants to retrieve the phonological neighborhood density (substitutions, additions, and deletions), the identity of the neighbors, word frequency, the zRT data for talker F2, and her sound files for these words. The user checks the relevant boxes, clicks download and will obtain a zip file that contains the wav files stored in a Sound Files folder together with the requested data in a csv file. Figure 10 shows the csv file as opened in Microsoft Excel for the example described.

For the words *abacus* and *zucchini*, there are no phonological neighbors, which is reflected in the NULL response for the iden_phono_n_sad field. Besides the mean zRT values (f2_ldt_zrt_m) for talker F2 requested by the user, the database will also automatically provide the *SD*s (f2_ldt_zrt_sd) and number of observations (f2_ldt_zrt_n) that the mean value is derived from. The database will automatically provide *SD*s and *n*s, where available, for all properties when the associated means are requested.

Downloads

In this section, users can download all sound files for words and nonwords, and all trial-level auditory lexical decision data for the six talkers. Figure 11 shows the trial-level data structure. A text file variables.txt describing the various fields is included in the zip file.

Table 17   zRT linear mixed model estimates for fixed and random effects across talkers

Fixed effects	F1	F2	F3	M1	M2	M3
Intercept	−0.05	−0.38***	−0.50***	−0.68***	0.01	−0.29***
Duration	0.002***	0.002***	0.002***	0.002***	0.002***	0.002***
PCA	−0.06***	−0.02*	−0.04**	−0.05***	−0.01	−0.04***
Number of morphemes	0.04***	0.05***	0.05***	0.05***	0.05***	0.06***
Phonological neighborhood density (substitution, addition, and deletion)	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
LgSUBTLCD	−0.08***	−0.07***	−0.08***	−0.06***	−0.1***	−0.08***
Prevalence	−0.12***	−0.15***	−0.15***	−0.15***	−0.15***	−0.12***
NUSFAM	−0.19***	−0.17***	−0.16***	−0.17***	−0.17***	−0.17***

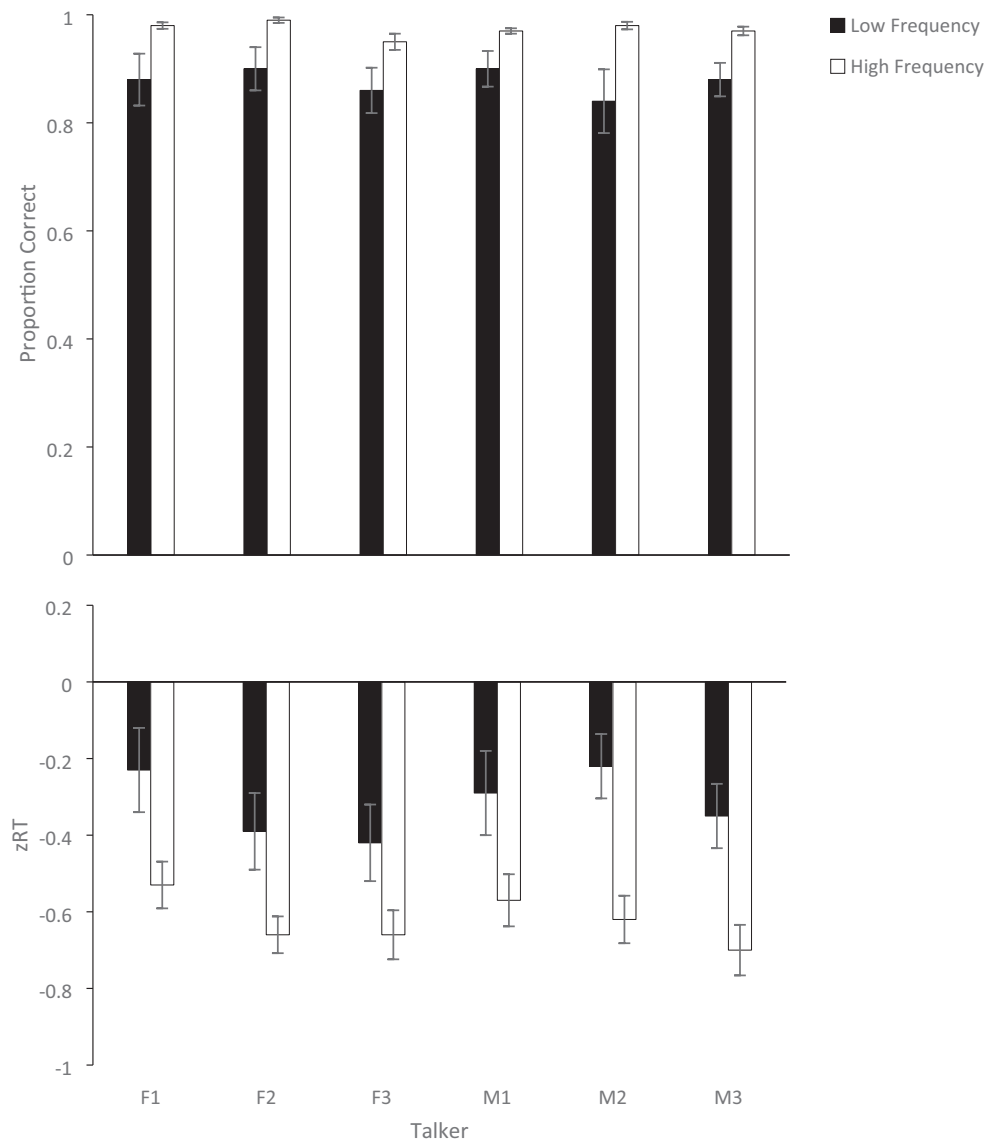* $p < .05$, ** $p < .01$, *** $p < .001$

Fig. 2 Accuracy (*top panel*) and latency (*bottom panel*) for virtual experiment of Taft and Hambly ([1986](#)) Experiment 4
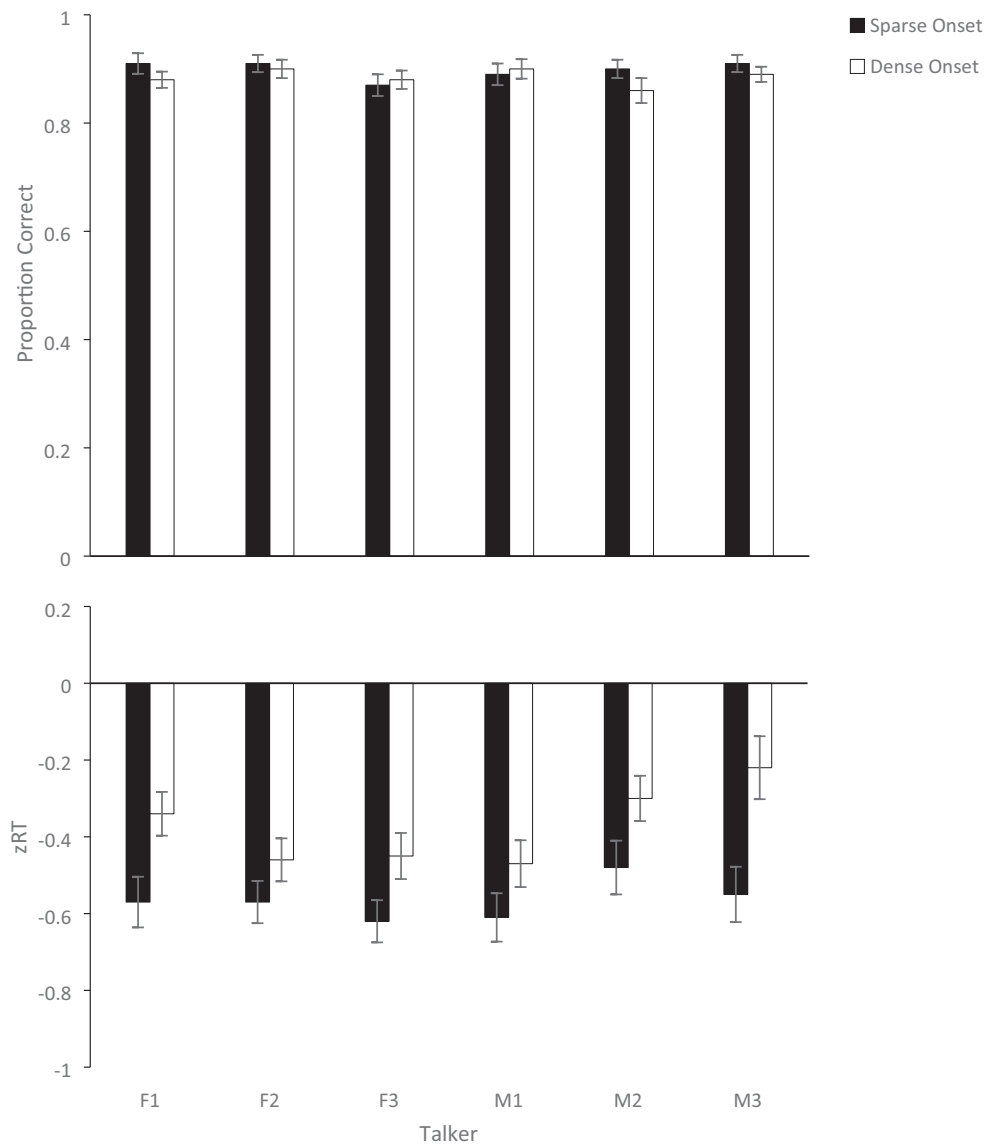
Fig. 3 Accuracy (*top panel*) and latency (*bottom panel*) for virtual experiment of Vitevitch (2002) Experiment 2
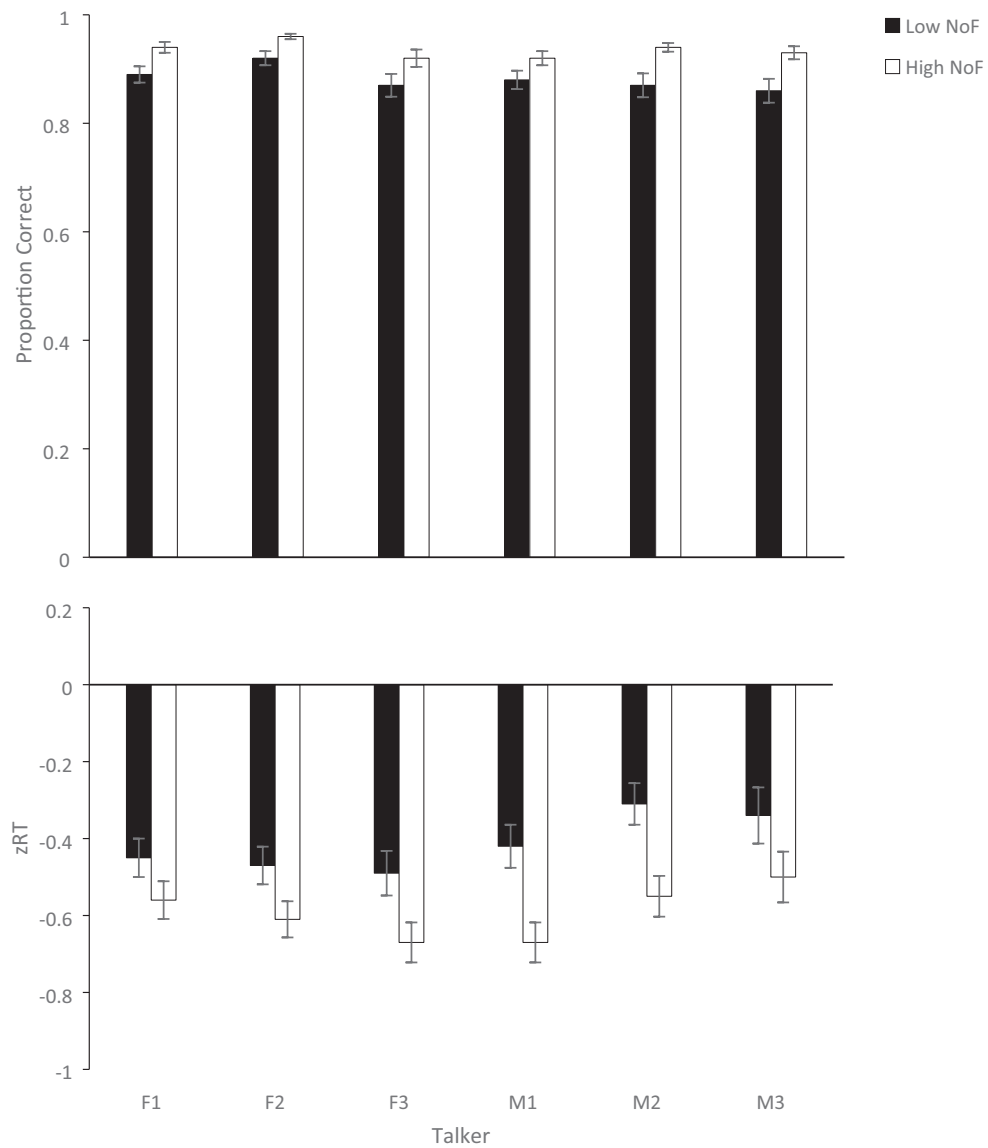
Fig. 4 Accuracy (*top panel*) and latency (*bottom panel*) for virtual experiment of Sajin and Connine (2014) Experiment 1
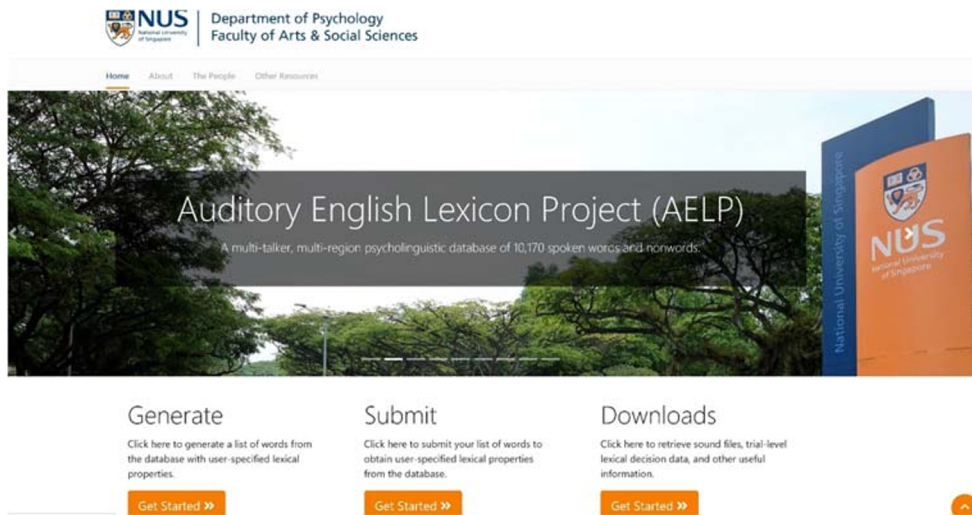

Fig. 5 Home page of the AELP website

Generate

Select the data type and dialect, then specify the metrics and variables required, along with the range of values. You can also specify whether you want the behavioural data and sound files.

1. Select Data:
○ Words only ● Words and Nonwords
☑ Include IPA pronunciations (nonwords only) ❶

2. Select Metrics:
Select Dialect:
○ American (US) ● British (UK)

Orthographic
☐ Word Length ❶ ⟶ Min 2 — 2 ▭ 17 — 17 Max

☐ Neighbourhood Density ❶
☐ Substitutions only ❶ ⟶ Min 0 — 0 ▭ 25 — 25 Max

☐ Substitutions, additions and deletions ❶ ⟶ Min 0 — 0 ▭ 31 — 31 Max

☐ Neighbourhood Frequency ❶
☐ Substitutions only ❶ ⟶ Min 0.1505 — 0.1505 ▭ 6.629 — 6.629 Max

☐ Substitutions, additions and deletions ❶ ⟶ Min 0.1505 — 0.1505 ▭ 6.0575 — 6.0575 Max

☐ Identity of Neighbours ❶
☐ Substitutions only ❶

☐ Substitutions, additions and deletions ❶

☐ Orthographic Spread ❶ ⟶ Min 0 — 0 ▭ 5 — 5 Max

☐ Orthographic Uniqueness Point ❶ ⟶ Min 2 — 2 ▭ 15 — 15 Max

☐ Orthographic Levenshtein Distance-20 (OLD-20) ❶ ⟶ Min 1 — 1 ▭ 8.55 — 8.55 Max

☐ Orthographic C Coefficient ❶ ⟶ Min 0 — 0 ▭ 1 — 1 Max

☐ Summed Bigram Frequency ❶ ⟶ Min 0.0001 — 0.0001 ▭ 0.6295 — 0.6295 Max

Phonological
☐ Number of Phonemes ⟶ Min 1 — 1 ▭ 17 — 17 Max

☑ Number of Syllables ⟶ Min 5 — 1 5 6 7 — 6 Max

☐ Neighbourhood Density ❶
☐ Substitutions only ❶ ⟶ Min 0 — 0 ▭ 36 — 36 Max

☐ Substitutions, additions and deletions ❶ ⟶ Min 0 — 0 ▭ 40 — 40 Max

[Reset Form]
[Download] ghbourhood Frequency ❶
Substitutions only ❶ ⟶ Min 0.1505 — 0.1505 ▭ 9.005 — 9.005 Max

Use me to avoid scrolling to the bottom!
☐ Substitutions, additions, and deletions ❶ ⟶ Min 0.1505 — 0.1505 ▭ 9.6287 — 9.6287 Max

Fig. 6 Generating a list of items based on user-specified properties

	A	B	C	D	E
1	word_uk	n_syll	nonword	nonword_ipa	n_syll_nw
2	abbreviation	5	abbreriotion	əbriːrɪˈəʊʃ(ə)n	5
3	abnormality	5	abnorcaloty	abnɒˈkaləʊti	5
4	abolitionist	5	abofitialist	əbɒˈfɪtɪəlɪst	5
5	abominable	5	abomogable	əˈbɒməgəb(ə)l	5
6	accommodation	5	accomrogation	əkɒmrəʊˈgeɪʃ(ə)n	5
7	accompaniment	5	accombarimics	əˈkʌmb(ə)rɪmɪks	5
8	actuality	5	actuority	aktʃjʊˈɒrɪti	5
9	administrative	5	adminustrotion	adˈmɪnʊstrəʊʃən	5
10	administrator	5	adminustrotar	adˈmɪnʊstrəʊtɑː	5
11	agricultural	5	agrilipteral	agrɪˈlɪptərəl	5
12	alliteration	5	allitenative	əlɪtəˈneɪtɪv	5
13	amateurism	5	amacourism	ˈaməkəʊrɪz(ə)m	5
14	Americanise	5	amerilanize	əˈmɛrɪl(ə)nʌɪz	5
15	amphitheatre	5	amphitreaten	ˈamfɪtriːeɪtən	5
16	amplification	5	amplirization	ˌamplɪrʌɪˈzeɪʃ(ə)n	5
17	anniversary	5	annisermagy	anɪˈsəːməgi	5
18	anonymously	5	anonygously	əˈnɒnɪgəsli	5
19	anthropology	5	anthrocalomy	ˌanθrəʊˈkaləʊmi	5
20	antibiotic	5	antifiuter	ˌantɪfɪuːˈtə	5
21	anticipated	5	anticimated	anˈtɪsɪmeɪtɪd	5
22	antipersonnel	5	anticermennic	antɪsəːmɛˈnɪk	5
23	apologetic	5	apolumetic	əpɒlʊˈmɛtɪk	5
24	appendicitis	5	appenmiciton	əˌpɛnmɪˈsɪtən	5
25	appropriation	5	approchiation	əˌprəʊtʃɪˈeɪʃ(ə)n	5
26	archaeology	5	archaealomy	ˌɑːkiːˈɒləmi	5
27	architectural	5	architullerly	ɑːkɪˈtuːləːli	5
28	argumentation	5	argucenraters	ˌɑːgjʊsɛnˈreɪtəs	5
29	argumentative	5	argucengators	ɑːgjʊˈsɛngətɔːz	5
30	aristocratic	5	aristustatic	ˌarɪstʊˈstatɪk	5
31	arithmetical	5	arithramical	ˌarɪθˈramɪk(ə)l	5
32	arithmetician	5	arithtenician	əˌrɪθtəˈnɪʃ(ə)n	5
33	articulation	5	articonation	ɑːˌtɪkəʊˈneɪtɪən	5

aelp_data ⊕

Fig. 7 List of words, the yoked nonwords (and IPA transcriptions), and number of syllables

Fig. 8 Submitting a list of items based on user specified properties (data and metrics)

Lexico-semantic Variables

☐ Number of Morphemes ❶
☑ Word Frequency (LgSUBTLEX) ❶
☐ Contextual Diversity (LgSUBTLCD) ❶
☐ Familiarity (FAM) ❶
   ☐ NUS FAM ❶
   ☐ Hoosier FAM ❶
☐ Word Prevalence ❶
   ☐ Prevalence
   ☐ Percentage Known
☐ Concreteness Rating ❶
☐ Age of Acquisition (AoA) ❶
☐ Semantic Neighbourhood Density (SND) ❶
   ☐ Neighbour Count (Ncount) ❶
   ☐ Average Radius of Co-occurrence (ARC) ❶
☐ Affective Ratings ❶
   ☐ Valence ❶
   ☐ Arousal ❶
   ☐ Dominance ❶

3. Select Behavioural Data:

Talker-level Estimates

For these talkers:

| ☐ Singaporean, Female (F1) | ☑ American, Female (F2) | ☐ British, Female (F3) |
| ☐ Singaporean, Male (M1) | ☐ American, Male (M2) | ☐ British, Male (M3) |

☐ Intelligibility Rate (ID) ❶
☐ Confidence Rating ❶
☐ LDT Accuracy ❶
☐ LDT Reaction Time (RT) ❶
☑ LDT z-score Reaction Time (zRT) ❶
☐ LDT z-score Reaction Time minus Duration (zRTmDur) ❶

Dialect-level Estimates

For these dialects:

| ☐ Singapore English (SG) | ☐ American English (US) | ☐ British English (UK) |

☐ LDT Accuracy ❶
☐ LDT Reaction Time (RT) ❶
☐ LDT z-score Reaction Time (zRT) ❶
☐ LDT z-score Reaction Time minus Duration (zRTmDur) ❶

4. Select Sound File Duration:

Talker-level Estimates

| ☐ Singaporean, Female (F1) | ☐ American, Female (F2) | ☐ British, Female (F3) |
| ☐ Singaporean, Male (M1) | ☐ American, Male (M2) | ☐ British, Male (M3) |

Dialect-level Estimates

| ☐ Singapore English (SG) | ☐ American English (US) | ☐ British English (UK) |

5. Select Sound Files:

☐ I do not need sound files

| ☐ Singaporean, Female (F1) | ☑ American, Female (F2) | ☐ British, Female (F3) |
| ☐ Singaporean, Male (M1) | ☐ American, Male (M2) | ☐ British, Male (M3) |

Download

Fig. 9 Submitting a list of items based on user specified properties (lexico-semantic variables, behavioral data, and sound files)

Also available for download are lists of homophones (in csv and Excel format) for American and British English, together with the frequency of each homophone response in the WordID data, and lists of participant

Fig. 10 Properties requested in the submit example

Fig. 11 Trial-level auditory lexical decision task data structure

characteristics for the ALD megastudy; these are described in Appendix 4. There is also a list of American and British spelling equivalents (e.g. *authorization-authorisation*, *color-colour*) so that users can input the correct spelling depending on which dialect they choose on the website.

Conclusions

The AELP is the largest multi-talker database of spoken words *and* nonwords to date, and will complement other single-talker large databases such as BALDEY, MEGALEX, and MALD. It is the first multi-*region* spoken word database that we are aware of, with speakers from three dialects of English. It is also the first to provide intelligibility norms for all tokens, which will be useful for researchers to decide whether the tokens can be used based on their requirements. The behavioral data replicates the classic findings of word frequency facilitation and word-form similarity inhibition for auditory

lexical decision, and also for word identification, which has not been shown before. It has further demonstrated facilitation effects for newer variables such as prevalence that has not been tested in the spoken domain. We hope that researchers will find the database useful.

Open Practices Statement The data and materials are available at https://inetapps.nus.edu.sg/aelp/ (https://doi.org/10.25542/5cvf-vv50). None of the data reported here were preregistered in an independent, institutional registry.

Supplemental Material There are three supplemental documents appended to this paper. S1 is the annotated Python script that was used to retrieve the phonological transcriptions from Oxford Dictionaries using their API; this is correct as of December 2019. S2 is an Excel file containing the base dataset of 26,604 words used for computing the structural properties of the AELP words. S3 contains the correlation matrices of all structural, lexico-semantic variables, and behavioural data for words and nonwords in an Excel file.

Appendix 1 Familiarity rating scale

Score	Label
1	I have never seen the word before.
2	I think that I might have seen the word somewhere before.
3	I am somewhat sure that I have seen the word before, but am not certain.
4	I have definitely seen the word before, but I don't know its meaning.
5	I am certain that I have seen the word before, but only have a vague idea about its meaning.
6	I think I might know the meaning of the word, but am not certain that the meaning I know is correct.
7	I recognize the word and am confident that I know at least one meaning.

Appendix 2 Descriptive statistics of database properties

Table 18   Structural properties

Property	Database field	American English					British English				
		N	Min	Max	M	SD	N	Min	Max	M	SD
Words											
<u>Orthographic</u>											
Word length	word_length	10170	2	17	7.20	2.44	10170	2	17	7.21	2.44
Neighborhood density											
Substitutions only	ortho_n_dens_s	10170	0	25	1.82	3.55	10170	0	25	1.82	3.53
Substitutions, additions, and deletions	ortho_n_dens_sad	10170	0	31	2.47	4.45	10170	0	31	2.46	4.43
Neighborhood frequency											
Substitutions only	ortho_n_freq_s_m	4204	0.15	5.83	2.21	0.74	4261	0.45	6.63	2.84	0.78
Substitutions, additions, and deletions	ortho_n_freq_sad_m	5110	0.15	5.47	2.19	0.74	5178	0.45	6.06	2.81	0.78
Identity of neighbors											
Substitutions only	iden_ortho_n_s	List of neighbors based on single letter substitution									
Substitutions, additions, and deletions	iden_ortho_n_sad	List of neighbors based on single letter substitution, addition, and deletion									
Orthographic spread	ortho_spread	10170	0	5	0.80	1.14	10170	0	5	0.80	1.14
Orthographic uniqueness point	ortho_upoint	10170	2	15	6.60	2.04	10170	2	15	6.61	2.04
Orthographic Levenshtein distance (OLD-20)	old20_m	10170	1.00	8.55	2.71	1.01	10170	1.00	8.55	2.71	1.01
Orthographic C coefficient	ortho_c_coeff	3359	0	1	.30	.28	3377	0	1	.31	.29
Summed bigram frequency	sum_bigram	10170	.00	.63	.07	.06	10170	.00	.60	.07	.06
<u>Phonological</u>											
Number of phonemes	n_phon	10170	1	17	6.42	2.45	10170	1	17	6.17	2.40
Number of syllables	n_syll	10170	1	7	2.41	1.16	10170	1	7	2.39	1.12
Neighborhood density											
Substitutions only	phono_n_dens_s	10170	0	35	3.24	5.74	10170	0	36	3.45	6.18
Substitutions, additions, and deletions	phono_n_dens_sad	10170	0	39	4.15	6.87	10170	0	40	4.39	7.36
Neighborhood frequency											
Substitutions only	phono_n_freq_s_m	5043	0.15	7.83	2.29	0.82	5043	0.48	9.01	2.96	0.91
Substitutions, additions, and deletions	phono_n_freq_sad_m	5819	0.15	6.94	2.27	0.80	5910	0.24	9.63	2.94	0.89
Identity of neighbors											
Substitutions only	iden_phono_n_s	List of neighbors based on single phoneme substitution									
Substitutions, additions, and deletions	iden_phono_n_sad	List of neighbors based on single phoneme substitution, addition, and deletion									
Phonological spread	phono_spread	10170	0	6	1.04	1.24	10170	0	5	1.04	1.24
Phonological uniqueness point	phono_upoint	10170	2	15	5.78	1.85	10170	2	14	5.61	1.85

Table 18 (continued)

Property	Database field	American English					British English				
		N	*Min*	*Max*	*M*	*SD*	*N*	*Min*	*Max*	*M*	*SD*
Phonological Levenshtein distance (PLD-20)	pld20_m	10170	1.00	7.80	2.47	1.00	10170	1.00	8.00	2.47	1.01
Phonological C coefficient	phono_c_coeff	4300	.00	1.00	.30	.25	4284	.00	1.00	.29	.24
Summed biphone frequency	sum_biphone	10167	.00	.67	.06	.06	10165	.00	.64	.05	.05
Phonographic											
Neighborhood density											
Substitutions only	pgraph_n_dens_s	10170	0	19	1.19	2.70	10170	0	19	1.22	2.73
Substitutions, additions, and deletions	pgraph_n_dens_sad	10170	0	24	1.69	3.45	10170	0	24	1.73	3.50
Neighborhood frequency											
Substitutions only	pgraph_n_freq_s_m	3154	0.30	5.52	2.29	0.76	3215	0.48	6.63	2.92	0.77
Substitutions, additions, and deletions	pgraph_n_freq_sad_m	3973	0.15	5.52	2.26	0.75	4068	0.24	5.87	2.89	0.77
Identity of neighbors											
Substitutions only	iden_pgraph_n_s	List of neighbors based on single letter and single phoneme substitution									
Substitutions, additions, and deletions	iden_pgraph_n_sad	List of neighbors based on single letter and single phoneme substitution, addition, and deletion									
Phonographic C coefficient	pgraph_c_coeff	2553	.00	1.00	.34	.31	2626	.00	1.00	.32	.30
Nonwords											
Phonological											
IPA transcription	nonword_ipa										
Number of phonemes	n_phon_nw	10170	2	17	6.51	2.43	10170	2	17	6.24	2.37
Number of syllables	n_syll_nw	10170	1	8	2.45	1.18	10170	1	7	2.41	1.13
Neighborhood density											
Substitutions only	phono_n_dens_s_nw	10170	0	31	1.92	4.04	10170	0	31	2.12	4.34
Substitutions, additions, and deletions	phono_n_dens_sad_nw	10170	0	39	2.27	4.66	10170	0	38	2.47	4.91
Neighborhood frequency											
Substitutions only	phono_n_freq_s_m_nw	3543	0.30	8.53	2.34	0.75	3735	0.48	7.20	3.00	0.81
Substitutions, additions, and deletions	phono_n_freq_sad_m_nw	3802	0.30	8.53	2.39	0.76	3991	0.48	8.29	3.05	0.81
Identity of neighbors											
Substitutions only	iden_phono_n_s_nw	List of word neighbors based on single phoneme substitution									
Substitutions, additions, and deletions	iden_phono_n_sad_nw	List of word neighbors based on single phoneme substitution, addition, and deletion									
Phonological spread	phono_spread_nw	10170	0	5	0.68	1.07	10170	0	5	0.72	1.08
Phonological uniqueness point	phono_upoint_nw	10170	2	11	4.27	1.06	10170	1	11	4.22	1.09
Phonological Levenshtein distance (PLD-20)	pld20_m_nw	10170	1.00	9.90	2.99	1.33	10170	1.00	9.65	2.88	1.25
Phonological C coefficient	phono_c_coeff_nw	2859	.00	1.00	.41	.29	3016	.00	1.00	.41	.28
Summed biphone frequency	sum_biphone_nw	10170	.00	.40	.04	.05	10170	.00	.43	.03	.04

Table 19 Lexico-semantic properties

Property	Database field	N	Min	Max	M	SD
Number of morphemes	n_morph	10170	1	5	1.51	0.66
Part of speech (POS): adjective, determiner, noun, pronoun, adverb, interjection, number, verb, conjunction, name, preposition	pos					
Word frequency	lgsubtlwf	9949	0.30	5.47	2.16	0.87
		9968	0.48	5.87	2.80	0.93
Contextual diversity	lgsubtlcd	9949	0.30	3.92	2.01	0.79
		9968	0.48	4.64	2.62	0.85
Familiarity (FAM)						
NUS FAM	nusfam_m	10170	2.17	7.00	6.63	0.63
Hoosier FAM	hoosierfam	9803	1.00	7.00	6.81	0.30
Word prevalence						
Prevalence	prevalence	10012	-0.52	2.58	2.17	0.33
Percentage known	pknown	10012	.30	1.00	.98	.03
Concreteness rating	concrete_m	9900	1.07	5.00	3.28	1.06
Age of acquisition (AoA)	aoa_m	9392	1.89	16.38	9.06	2.66
Semantic neighborhood density (SND)						
Neighbour count (Ncount)	snd_ncount	9959	0	9901	1927.61	2845.60
Average radius of co-occurrence (ARC)	snd_arc	9959	.11	.70	.50	.14
Affective ratings						
Valence	aff_val_m	7612	1.30	8.53	5.18	1.29
Arousal	aff_arou_m	7612	1.60	7.79	4.19	0.91
Dominance	aff_dom_m	7612	2.14	7.90	5.27	0.95

For word frequency and contextual diversity, the top and bottom numbers represent the values from the US and UK databases, respectively

Appendix 3 Descriptive statistics of lexical decision measures and token durations

Table 20 Talker-level estimates

Talker	Database field	Min	Max	M	SD
Words					
Accuracy					
F1	f1_ldt_acc	.04	1.00	.89	.15
F2	f2_ldt_acc	.06	1.00	.91	.13
F3	f3_ldt_acc	.04	1.00	.88	.16
M1	m1_ldt_acc	.07	1.00	.89	.15
M2	m2_ldt_acc	.03	1.00	.89	.15
M3	m3_ldt_acc	.03	1.00	.89	.15
Raw reaction time (mean)					
F1	f1_ldt_rt_m	604.33	1478.50	902.53	113.25
F2	f2_ldt_rt_m	683.03	1812.50	1018.28	121.90
F3	f3_ldt_rt_m	608.79	1709.60	975.03	135.84
M1	m1_ldt_rt_m	706.63	1730.63	1070.83	137.77
M2	m2_ldt_rt_m	667.32	1503.69	982.87	113.84
M3	m3_ldt_rt_m	632.19	1528.58	968.01	126.68
z-scored reaction time (mean)					
F1	f1_ldt_zrt_m	−1.41	2.05	−0.25	0.45
F2	f2_ldt_zrt_m	−1.54	1.67	−0.29	0.44
F3	f3_ldt_zrt_m	−1.59	2.33	−0.23	0.52

Table 20 (continued)

Talker	Database field	Min	Max	M	SD
M1	m1_ldt_zrt_m	−1.76	1.79	−0.18	0.51
M2	m2_ldt_zrt_m	−1.50	1.59	−0.19	0.46
M3	m3_ldt_zrt_m	−1.66	1.72	−0.23	0.51
z-scored [reaction time minus token duration] (mean)					
F1	f1_ldt_zrtmdur_m	−1.79	2.01	−0.18	0.48
F2	f2_ldt_zrtmdur_m	−2.16	1.76	−0.21	0.44
F3	f3_ldt_zrtmdur_m	−1.84	2.79	−0.12	0.51
M1	m1_ldt_zrtmdur_m	−2.18	2.18	−0.15	0.51
M2	m2_ldt_zrtmdur_m	−1.81	1.99	−0.15	0.51
M3	m3_ldt_zrtmdur_m	−2.47	2.05	−0.14	0.54
Token duration (ms)					
F1	f1_duration	284.35	1328.30	689.65	143.78
F2	f2_duration	423.83	1478.50	838.39	162.60
F3	f3_duration	338.59	1682.59	789.91	187.82
M1	m1_duration	362.68	1899.84	937.89	202.64
M2	m2_duration	344.63	1522.29	820.57	165.11
M3	m3_duration	347.21	1555.44	837.23	186.50
Nonwords					
Accuracy					
F1	f1_ldt_acc_nw	.08	1.00	.91	.10
F2	f2_ldt_acc_nw	.00	1.00	.90	.12
F3	f3_ldt_acc_nw	.07	1.00	.92	.10
M1	m1_ldt_acc_nw	.14	1.00	.92	.10
M2	m2_ldt_acc_nw	.00	1.00	.92	.11
M3	m3_ldt_acc_nw	.10	1.00	.92	.10
Raw reaction time (mean)					
F1	f1_ldt_rt_m_nw	747.73	1452.68	996.77	88.03
F2	f2_ldt_rt_m_nw	754.00	1654.25	1145.08	100.73
F3	f3_ldt_rt_m_nw	758.44	1724.00	1061.56	111.01
M1	m1_ldt_rt_m_nw	799.68	1617.19	1128.62	111.50
M2	m2_ldt_rt_m_nw	701.00	1601.22	1041.91	89.64
M3	m3_ldt_rt_m_nw	748.21	1497.67	1046.19	99.12
z-scored reaction time (mean)					
F1	f1_ldt_zrt_m_nw	−0.89	1.49	0.10	0.34
F2	f2_ldt_zrt_m_nw	−1.14	1.67	0.14	0.37
F3	f3_ldt_zrt_m_nw	−1.15	1.78	0.09	0.43
M1	m1_ldt_zrt_m_nw	−1.23	1.69	0.03	0.42
M2	m2_ldt_zrt_m_nw	−1.40	1.56	0.03	0.37
M3	m3_ldt_zrt_m_nw	−1.10	1.84	0.08	0.40
z-scored [reaction time minus token duration] (mean)					
F1	f1_ldt_zrtmdur_m_nw	−1.82	1.99	0.05	0.50
F2	f2_ldt_zrtmdur_m_nw	−2.14	2.51	0.08	0.48
F3	f3_ldt_zrtmdur_m_nw	−2.28	2.07	0.01	0.56
M1	m1_ldt_zrtmdur_m_nw	−2.46	2.59	0.04	0.55
M2	m2_ldt_zrtmdur_m_nw	−2.15	1.86	0.02	0.55
M3	m3_ldt_zrtmdur_m_ nw	−2.14	1.94	0.02	0.59
Token duration (ms)					
F1	f1_duration_nw	287.14	1313.79	718.92	141.61
F2	f2_duration_nw	437.35	1642.88	877.02	166.90

Table 20 (continued)

Talker	Database field	Min	Max	*M*	*SD*
F3	f3_duration_nw	351.81	1889.46	840.76	196.69
M1	m1_duration_nw	334.83	1786.03	943.79	206.02
M2	m2_duration_nw	362.47	1680.39	829.96	171.72
M3	m3_duration_nw	240.59	1595.56	869.78	188.11

Table 21 Dialect-level estimates

Dialect	Database field	Min	Max	*M*	*SD*
Words					
Accuracy					
Singapore English	sg_ldt_acc	.09	1.00	.89	.14
American English	us_ldt_acc	.08	1.00	.90	.13
British English	uk_ldt_acc	.04	1.00	.89	.15
Raw reaction time (mean)					
Singapore English	sg_ldt_rt_m	686.49	1529.88	989.30	113.44
American English	us_ldt_rt_m	693.00	1529.00	1000.04	108.87
British English	uk_ldt_rt_m	630.61	1498.71	970.47	120.65
z-scored reaction time (mean)					
Singapore English	sg_ldt_zrt_m	−1.50	1.55	−0.22	0.44
American English	us_ldt_zrt_m	−1.42	1.47	−0.24	0.42
British English	uk_ldt_zrt_m	−1.50	1.92	−0.23	0.48
z-scored [reaction time minus token duration] (mean)					
Singapore English	sg_ldt_zrtmdur_m	−1.63	1.40	−0.17	0.46
American English	us_ldt_zrtmdur_m	−1.69	1.51	−0.18	0.43
British English	uk_ldt_zrtmdur_m	−1.88	2.42	−0.13	0.49
Token duration (ms)					
Singapore English	sg_duration	362.22	1523.72	813.77	163.95
American English	us_duration	384.23	1475.33	829.48	151.48
British English	uk_duration	368.12	1559.24	813.57	176.61
Nonwords					
Accuracy					
Singapore English	sg_ldt_acc_nw	.26	1.00	.88	.10
American English	us_ldt_acc_nw	.12	1.00	.91	.10
British English	uk_ldt_acc_nw	.16	1.00	.88	.10
Raw reaction time (mean)					
Singapore English	sg_ldt_rt_m_nw	813.37	1542.55	1066.19	86.11
American English	us_ldt_rt_m_nw	806.63	1528.48	1091.96	82.94
British English	uk_ldt_rt_m_nw	801.86	1543.70	1053.44	93.51
z-scored reaction time (mean)					
Singapore English	sg_ldt_zrt_m_nw	−0.90	1.57	0.06	0.34
American English	us_ldt_zrt_m_nw	−0.99	1.42	0.08	0.33
British English	uk_ldt_zrt_m_nw	−0.93	1.60	0.08	0.37
z-scored [reaction time minus token duration] (mean)					
Singapore English	sg_ldt_zrtmdur_m_nw	−1.81	2.02	0.04	0.49
American English	us_ldt_zrtmdur_m_nw	−1.75	1.64	0.05	0.47
British English	uk_ldt_zrtmdur_m_nw	−2.00	1.94	0.01	0.54
Token duration (ms)					
Singapore English	sg_duration_nw	394.52	1524.10	831.35	164.17
American English	us_duration_nw	428.19	1661.63	853.49	156.79
British English	uk_duration_nw	379.89	1639.09	855.27	181.06

Dialect token duration estimates is the average of the male and female tokens

Appendix 4. Other lists in the downloads section of the website

Figure 12 shows the homophones for American and British English in Excel format. Some words have the same homophones in both dialects, e.g. *altar* is a word in the AELP database, and the corresponding homophone for both dialects is *alter*. The talker ID columns indicate the number of responses recorded for each homophone in the WordID sessions for each talker. Some words have a homophone in one dialect but not the other. For example, *aunt* is a homophone for *ant* in American but not British English. This list is useful for users who need to estimate what homophone is likely to be perceived by listeners for a particular target word.

Fig. 12 List of homophones

Table 22 summarizes the participant characteristic fields and value labels for the ALD megastudy.

Table 22 Participant characteristics for auditory lexical decision megastudy

Variable Name	Description	Values
Gender		1 Male; 2 Female
Hand	Handedness	1 Left; 2 Right; 3 Ambidextrous
Ghent[1]	Ghent vocabulary score	
Vocab	Vocabulary score from Levy et al. (2017)	
Spelling	Spelling score from Burt & Tate (2002)	
Country/Region_x[2]	Countries or regions lived in	1 Singapore; 2 Australia; 3 China; 4 England; 5 Germany; 6 Hong Kong SAR; 7 India; 8 Indonesia; 9 Malaysia; 10 Myanmar; 11 Taiwan; 12 USA; 13 Burma; 14 Canada; 15 Laos; 16 Philippines
CRx_Age	Age when started living in country/region	
CRx_Year	Number of years lived in country/region	
L1	First language	1 English; 2 Chinese; 3 Malay; 4 Tamil; 5 Chinese dialect; 6 Indian dialect; 7 Japanese; 8 Korean; 9 German; 10 French; 11 Spanish; 12 Russian; 13 Thai; 14 Vietnamese; 15 Bahasa Indonesia; 16 Khmer; 17 Arabic; 18 Tagalog; 19 Swedish; 20 Javanese; 21 Burmese
L2	Second language	
L3	Third language	
L4	Fourth language	
Lx_AOA	Age of acquisition	
Lx_L	Listening proficiency (0–7)	Self-rated higher values represent greater proficiency
Lx_R	Reading proficiency (0–7)	
Lx_S	Speaking proficiency (0–7)	
Lx_W	Writing proficiency (0–7)	

1. To screen out non-native English speakers, we had asked prospective participants to report their scores for the Ghent University Word Test at http://vocabulary.ugent.be/ (retrieved 24 October 2019), adopting a minimum criterion of 40% to take part in the Word recognition phases. See also Brysbaert, et al. (2019).

2. "x" represents a number. For example, if the participant lived in more than one region/country, CR1_Age and CR2_Age would represent the age he or she started living in those countries/regions. Likewise, if they know more than one language, L1_L and L2_L would represent their self-reported listening proficiency for the first and second languages, respectively.

References

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighbourhood conflicts. *Psychonomic Bulletin & Review*, *4*, 439–461.

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.

Balota, D., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (A travel update). In M. Traxler & M. A. Gernsbacher (Eds.) Handbook of psycholinguistics *(*2nd*)* (pp. 285–375). Amsterdam: Academic Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … , Treiman, R. (2007). The English Lexicon Project. *Behaviour Research Methods*, *39*, 445–459.

Balota, D. A., Yap, M. J., Hutchison, K.A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In James S. Adelman (Ed). Visual word recognition (pp. 90–115). Hove: Psychology Press.

Bates, D., Maechler, M., Bolker, M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi:https://doi.org/10.18637/jss.v067.i01

The British National Corpus, version 3 (BNC XML Edition) (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. Retrieved from: http://www.natcorp.ox.ac.uk/

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*, 467–479.

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. *Behaviour Research Methods*, *41*, 977–990.

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behaviour Research Methods*, *46*, 904–911.

Burt, J. S., & Tate, H. (2002). Does a reading lexicon provide orthographic representations for spelling? *Journal of Memory and Language, 46*(3), 518–543.

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behaviour Research Methods, Instruments, & Computers*, *36*, 384–387.

Courrieu, P., & Rey, A. (2011). Missing data imputation and corrected statistics for large-scale behavioral databases. *Behavior Research Methods*, *43*, 310–330. doi:https://doi.org/10.3758/s13428-011-0071-2.

Crowder, R. G. (1971). The sound of vowels and consonants in immediate memory. *Journal of Verbal Learning and Verbal Behavior*, *10*, 587–596.

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics*, *5*, 365–373.

Cutler, A. (1981). Making up materials is a confounded nuisance: or will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*, 65–70.

Dahan, D., & Magnuson, J. S. (2006). Spoken word recognition. In M. Traxler & M. A. Gernsbacher (Eds.) Handbook of psycholinguistics (2nd*)* (pp. 249–283). Amsterdam: Academic Press.

Deterding, D. (2007). Singapore English. Edinburgh: Edinburgh University Press.

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology*, *68*, 1469–1488.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777–799.

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., …, Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition. *Behaviour Research Methods*, *50*, 1285–1307.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Philadelphia: Linguistic Data Consortium.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*, 256–281.

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 40–53.

Goh, W. D., Suárez, L., Yap, M. J., & Tan, S. H. (2009). Distributional analyses in auditory lexical decision: Neighbourhood density and word frequency effects. *Psychonomic Bulletin & Review*, *16*, 882–887.

Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M. R., & Tan, L. C. (2016). Semantic richness effects in spoken word recognition: A lexical decision and semantic categorisation megastudy. *Frontiers in Psychology*, *7*, 976.

Goldinger, S. D. (1996a). Auditory lexical decision. *Language and Cognitive Processes*, *11*, 559–567.

Goldinger, S. D. (1996b). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 252–279.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, *37*, 158–66.

IEEE. (1969). *IEEE recommended practice for speech quality measurements* (IEEE Report No. 297).

Johnson, N. F., & Pugh, K. R. (1994). A cohort model of visual word recognition. *Cognitive Psychology*, *26*, 240–346.

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

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*, 287–304.

Kreuz, R. J. (1987). The subjective familiarity of English homophones. *Memory & Cognition*, *15*, 154–168.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer Objects of Lme4 Package)*: R Package Version 2.0-6. Available online at: http://CRAN.R-project.org/package=lmerTest

Leimgruber, J. R. E. (2011). Singapore English. *Language and Linguistics Compass*, *5*, 47–62.

Levy, J., Bullinaria, J., & McCormick, S. (2017). Semantic vector evaluation and human performance on a new vocabulary MCQ test. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), Proceedings of the 39th Annual Conference of the Cognitive Science Society (pp. 2549–2554). Austin, TX: Cognitive Science Society.

Lim, L. C. L., & Goh, W. D. (2019). False recognition modality effects in short-term memory: Reversing the auditory advantage. *Cognition*, *193*, https://doi.org/10.1016/j.cognition.2019.104008

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, *89*, 874–886.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, *39*, 155–158.

Luce, P. A., & Pisoni, D. B. (1998). Recognising spoken words: The neighbourhood -activation model. *Ear & Hearing*, *19*, 1–36.

Makino, S., Abe, M., & Kido, K. (1988). An isolated spoken word database using CD-ROMs. *Journal of the Acoustical Society of America*, *84*, S218.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29–63.

McRae, K., Cree, G., Seidenberg, M., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behaviour Research Methods, Instruments, & Computers*, *37*, 547–559.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42–46.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). *Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words* (Research on Speech Perception Progress Report No. 10). Bloomington: Speech Research Laboratory, Department of Psychology, Indiana University.

Olszewska, J. M., Reuter-Lorenz, P. A., Munier, E., & Bendler, S. A. (2015). Misremembering what you see or hear: Dissociable effects of modality on short- and long-term false recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *41*, 1316–1325.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.

Oxford Dictionaries. (2019). Oxford University Press: Oxford, United Kingdom. Retrieved from: https://www.lexico.com

Peereman, R., & Content, A. (1997). Orthographic and phonological neighbourhoods in naming: Not all neighbours are equally influential in orthographic space. *Journal of Memory & Language*, *37*, 382–410.

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition*, *17*, 398–422.

Pexman, P. M. (2012). Meaning-based influences on visual word recognition. In J. S. Adelman (Ed.), Visual Word Recognition Volume 2 (pp. 24–43). Hove: Psychology Press.

Pisoni, D. B. (1996). Word identification in noise. *Language and Cognitive Processes*, *11*, 681–688.

Pisoni, D. B., & Lively, S. E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), Speech perception and linguistic experience (pp. 433–459). Baltimore: York Press.

R Core Team (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*, 385–407.

Rosenzweig, M. R., & Postman, L. (1957). Intelligibility as a function of frequency of usage. *Journal of Experimental Psychology*, *54*, 412–422.

Sajin, S. M., & Connine, C. M. (2014). Semantic richness: The role of semantic features in processing spoken words. *Journal of Memory and Language*, *70*, 13–35.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behaviour Research Methods*, *42*, 393–413.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime user's guide. Pittsburgh: Psychology Software Tools Inc.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). E-Prime 2.0 user's guide. Pittsburgh: Psychology Software Tools Inc.

Schock, J., Cortese, M. J., & Khanna, M. M. (2012). Imageability estimates for 3,000 disyllabic words. *Behaviour Research Methods*, *44*, 374–379.

Shipley, W.C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology*, *9*, 371–377.

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106*, 833–870.

Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbours. *Journal of Memory and Language*, *40*, 83–108.

Taft, M., & Hambly, G. (1986). Exploring the Cohort Model of spoken word recognition. *Cognition*, 22, 259–282.

Tay, M. W. J., & Gupta, A. F. (1983). Towards a description of Standard Singapore English. In R. B. Noss (ed.), Varieties of English in Southeast Asia (pp. 173–189). Singapore: SEAMEO Regional Language Centre.

Treisman, M. (1978). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior*, *17*, 37–59.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behaviour Research Methods*, *51*, 1187–1204.

Vitevitch, M. S. (2002). Influence of onset density on spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 270–278.

Vitevitch, M. S. (2007). The spread of the phonological neighborhood influences spoken word recognition. *Memory & Cognition, 35,* 166–175.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research, 51,* 408–422.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers*, *36*, 481–487.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behaviour Research Methods*, *45*, 1191–1207.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature, 393,* 409–410.

Yarkoni, T., Balota, D., & Yap, M. J. (2008). Moving beyond Coltheart's N: a new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*, 971–979.