Behavior Research Methods (2019) 51:1187–1204
https://doi.org/10.3758/s13428-018-1056-1

The Massive Auditory Lexical Decision (MALD) database

Benjamin V. Tucker[1] · Daniel Brenner[1] · D. Kyle Danielson[2] · Matthew C. Kelley[1] · Filip Nenadić[1] · Michelle Sims[1]

Published online: 18 June 2018

Abstract
The Massive Auditory Lexical Decision (MALD) database is an end-to-end, freely available auditory and production data set for speech and psycholinguistic research, providing time-aligned stimulus recordings for 26,793 words and 9592 pseudowords, and response data for 227,179 auditory lexical decisions from 231 unique monolingual English listeners. In addition to the experimental data, we provide many precompiled listener- and item-level descriptor variables. This data set makes it easy to explore responses, build and test theories, and compare a wide range of models. We present summary statistics and analyses.

Keywords Megastudy · Auditory lexical decision · Spoken word recognition

Introduction

In psycholinguistics, databases constructed using large-scale studies with responses to tens of thousands of words have become an important resource for researchers investigating many topics, particularly lexical processing and representation. In the present paper, we detail the design and construction of what we call the "Massive Auditory Lexical Decision" (MALD) database, which provides an auditory parallel to many of the existing large-scale databases that have been conducted almost exclusively in the visual domain.

One of the first data sets of this scale was the English Lexicon Project (Balota et al., 2007), which contains responses from visual lexical decision and naming tasks of North American English. More recently, similar studies have been completed for British English (Keuleers et al., 2012), Dutch (Keuleers et al. 2010, 2015), French (Ferrand et al., 2010), and Cantonese (Tse et al., 2016), as well as a database of lexical decision and naming across the lifespan for German (Schröter & Schroeder, 2017). These large-scale studies or "megastudies" (Seidenberg & Waters, 1989) have several important advantages, including statistical power, minimization of strategic effects, comprehensiveness, and multi-functionality, as well as complementing and validating a wide range of traditional, smaller factorial experiments (Balota et al., 2012; Keuleers & Balota, 2015).

Recently, the MEGALEX database has been released, which investigates both visual and auditory recognition of spoken French (Ferrand et al., 2017) helping to move the megastudy research into the auditory domain. Preceding this work, several smaller-scale auditory databases have been produced (Luce & Pisoni, 1998; Smits et al., 2003; Warner et al., 2014; Ernestus & Cutler, 2015), which investigate the processing of spoken language. Even though speech is phylogenetically and ontogenetically prior to reading, and despite the fact that even in modernized literate societies most daily interactions likely occur in the form of speech rather than reading or writing, megastudies are still largely restricted to the visual domain. In the remainder of this section, we briefly review the advantages of megastudies and we discuss in more detail the specific motivation for the creation of the present auditory lexical decision data set.

Why megastudies?

Balota et al. (2012) and Keuleers and Balota (2015) provide summaries of the history and advantages of megastudies in psycholinguistics. In this section and the next, we give our primary motives for producing MALD, why a megastudy was warranted, and in particular why one is overdue for English in the auditory domain.

✉ Benjamin V. Tucker
bvtucker@ualberta.ca

[1] Department of Linguistics, University of Alberta, Edmonton, AB, Canada

[2] University of Toronto, Toronto, ON, Canada

Small targeted experiments often suffer from sampling difficulties. Smaller samples of items are more susceptible to sampling bias, not only of the random variety, but also due to the conditions imposed by the theoretical question itself and properties of the language. For example, if a researcher is investigating differences between closed- and open-class words, the number of items and their lengths and makeup will be constrained by the set of closed-class words existing in the language. In addition to these types of constraints, item selection can also be susceptible to the experimenters' unconscious selection decisions (Cutler, 1981; Forster, 2000). A large data set provides a theoretical testing ground in which the item selection is independent of the question and the experimenter. Naturally, such a data set is still subject to the limitations of the specific language, and in some cases may be even more limited than smaller data sets created for the purposes of a single study. These limitations often motivate the conjunction of large-scale studies with smaller, targeted studies.

In the role of complementary research, large databases enable the quick confirmation of experimental results with an independent sample of items, subjects, and responses, collected by researchers unrelated to, and uninfluenced by, the studied effects. Their size also affords the statistical power to detect small effects and reduces the possibility of type II error, which has been a serious recurring issue in the social sciences (Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Ioannidis, 2005; Maxwell et al., 2015). Many smaller studies would be buttressed considerably by the type of verification provided by larger data sets. All studies have limitations and biases, and the more corroborative evidence that can be gathered from a variety of sources, samples, and methods, the more persuasive their conclusions can be (Campbell, 1959; Shadish, 1993). A general-purpose large database can contribute in this capacity to a variety of studies.

Using data from a megastudy, a researcher can create a virtual experiment using a list of items that match a theoretical question and conduct the experiment with a statistical analysis (e.g., Kuperman (2015), Baayen et al. (2017), Brysbaert and New (2009), Brysbaert et al. (2016), and New et al. (2006)). That is, factorial designs are important for understanding a particular question, but augmenting them with data from megastudies or using data from a megastudy as a pilot before investing resources on a particular experiment is a useful and expedient research option. As previously noted, it also allows the researcher to avoid the influence of unconscious selection decisions (Cutler, 1981; Forster, 2000).

Another advantage of megastudies is that they allow for rapid theoretical development and complex modeling. Large databases of visual word recognition responses have been extremely useful in testing models of visual language processing, and offered new possibilities in computational modeling (Norris, 2013). For example, the English Lexicon Project and other visual megastudies have been used extensively by proponents of multiple models to gauge the importance of different lexical factors (e.g., New et al. 2006; Dufau et al. 2012; Yap et al. 2012, 2015; Mandera et al. 2017). In speech perception and comprehension, although certain attempts which we will describe below have been made, such large databases have not been readily available to researchers. Still, Shortlist B (Norris & McQueen, 2008) was developed on the basis of a large database of Dutch diphone perception (Smits et al., 2003). A more recent model, DIANA (ten Bosch et al. 2013, 2014, 2015a, b), was created based on a newly developed large data set, described in more detail below (Ernestus & Cutler, 2015).

Auditory vs. visual perception

While there has been a preponderance of megastudies in the visual modality, there is a general lack of these in the auditory modality. In this section we will discuss some of the reasons why this is the case, describe some of the differences between the modalities, and briefly describe the existing auditory data that we are aware of.

Compared to reading experiments, auditory experiments are labor-intensive. This may partly account for the general trend that the psycholinguistic aspects of auditory comprehension are under-researched compared to visual comprehension. Whereas it is relatively less time-consuming and more straightforward to create and control stimuli in a visual experiment, an experiment requiring the creation of auditory stimuli is often a complex multistep process requiring a great deal of human effort, as described below in the Methods section ("Methods"). Experiments in the two modalities are similar in the design, balancing, and selection of target words and foils (though in the auditory case one indexes a pronunciation dictionary so as to work in phones rather than orthographic glyphs). However, the two designs quickly diverge in the stimulus creation stage. In the visual case, this normally involves simple computer output to a standard font, while the auditory case typically demands recording, markup, and extraction of items, auditory and visual location of acoustic landmarks or relevant intervals, and audio normalization or other post-processing, all of which involve human labor. The labor-intensive nature of the setup for experiments in the auditory domain serves as a disincentive for many researchers to perform such work. However, as we will detail, work in the visual domain is no substitute for corresponding work in the auditory domain.

Speech unfolds over time, and even single words are not taken up or auditorily "glimpsed" in a single chunk, but are integrated over time by a running process (Mattys, 1997; Smits et al., 2003; Warner et al., 2014). This is

quite different from the processing for written words, which are ordinarily chunked in brief individual fixations of the eyes (Rayner & Clifton, 2009; Rayner et al., 2006; Radach & Kennedy, 2013). Another characteristic that sets reading apart from listening is the ability to re-access the information source, either by fixating it for a longer period, or by making a regressive eye movement backward in the text. In the auditory modality, listeners rarely have the option of replaying the sound. Consequently, the existing large visual processing studies (Balota et al. 2007; Ferrand et al. 2010; Keuleers et al. 2010, 2012) are inadequate for understanding how auditory language comprehension occurs.

It should not be surprising that the organization of an individual's knowledge about speech might differ from their knowledge about writing, or that the processes involved in perception of the two kinds of stimuli should differ. The psycholinguistic relevance of these differences can be seen when we consider the details of the findings in the two modalities. For example, the effect of neighborhood density varies depending on the modality and how neighborhood density is calculated (orthographic vs. phonological). In visual lexical decision, both orthographic and phonological neighborhood density is facilitative (Coltheart et al., 1977; Andrews, 1997; Yates et al., 2004), making the response latencies shorter in high-density neighborhoods. However, in the auditory modality, one of the effects changes: the effect of orthographic neighborhood density remains facilitative (Ziegler et al., 2003), but phonological neighborhood density is inhibitive (Luce & Pisoni, 1998; Vitevitch & Luce, 1998).

Conclusions about lexical processing based on reading often do not provide a satisfactory description of auditory comprehension, though some have claimed that models of visual lexical access should be appropriate for speech, accounting for obvious physical differences (Bradley & Forster, 1987). Even so, most models of lexical access have diverged such that there are different models across the modalities. Further, the visual signal cannot address the inherent variation in the spoken signal, within individual speakers, across speakers, and across dialects. While letter shapes are effectively invariant across words (for a given font), the acoustic realization of speech elements is decidedly variable.

Large databases in the auditory modality

As we noted previously, there are numerous benefits of creating large databases in the field of psycholinguistics, and there is also a relative lack of megastudies for auditory word recognition in comparison to visual studies, even though the two modalities cannot be equated. However, there are a few studies for auditory processing that can be considered megastudies.

In their experiments investigating the effects of neighborhood density and the neighborhood activation model, Luce and Pisoni (1998) tested 918 monosyllabic words of the form CVC. Ninety participants listened to one of three sublists (306 words each) with words presented in three different noise conditions (signal-to-noise ratios: +15, +5, and −5 dB) and were asked to identify the word by typing the word. An additional 30 participants performed an auditory lexical decision task in a separate experiment. In this case, words were again divided into three lists and each participant heard 306 words and 304 pseudowords. The resultant data set produced ten observations per word for all 918 words.

There have also been two studies that have examined Dutch and English diphone processing (Smits et al., 2003; Warner et al., 2014). In both of these experiments, participants were presented with gated fragments of diphones and were asked to identify them. In the Dutch experiment, there were 1170 diphones and in the English experiment there were 2288 diphones. The Dutch database has played a central role in the development of Shortlist B (Norris & McQueen, 2008).

Ernestus and Cutler (2015) published the 'Biggest Auditory Lexical Decision Experiment Yet', or BALDEY. This data set for Dutch contains auditory lexical decision responses from 20 participants who each responded to 2780 words and 2762 pseudowords resulting in 110,820 responses. The stimuli purposely sample a wide range of morphological complexity, including compound words.

Recently, Ferrand et al. (2017) produced the MEGALEX database. This database investigates comprehension of French and contains data from both the visual and auditory modalities, with a specific focus on comparing comprehension in the two modalities. The visual experiment contains 28,466 words and the same number of pseudowords, while the auditory experiment contains 17,876 words and the same number of pseudowords. In this experiment, as opposed to BALDEY, the authors used speech synthesis to create all of their stimuli. The MEGALEX and BALDEY data sets are the closest data sets in size and approach to the MALD data set described below.

In the remainder of the paper, we describe the design and creation of the MALD database, and summarize general properties of the items, listeners, and responses.

Methods

The goal of item selection for the project was to ensure generalization across the spoken English lexicon, ultimately including 26,793 words and 9592 pseudowords in the data set reported here.

Items

Words

The list of words for MALD was compiled to represent conversational speech with the final goal of a list of at least 25,000 words. To do this, we extracted all the unique word types (about 8000) in the Buckeye Corpus of Conversational Speech (Pitt et al., 2007), which provided a base of words biased toward conversational speech. The base word list was then augmented with words from COCA (Davies, 2009). We selected the first 25,000 words (ranked by frequency of occurrence) and merged these words with the base list. After removing specific entries, defined below, the base list was augmented with about 10,000 words from COCA. The word list was also augmented with a list of 1252 compound words extracted from CELEX (Baayen et al., 1995). The base list was then further augmented with about 9000 words randomly sampled from the English Lexicon Project (Balota et al., 2007).

During the compilation of the word list, we attempted to exclude items that were proper nouns, coordinating conjunctions, offensive words, days of the week, and letters of the alphabet. The resulting list contained a total of 28,510 words and includes mono- and multi-morphemic words, inflected and derived forms, function and content words, compound words, and all parts of speech apart from proper nouns and coordinating conjunctions. For pronunciation referencing, we used the CMU Pronouncing Dictionary V0.6 (Weide, 2005, 133,315 pronunciations of 123,656 unique headwords), augmented with 3726 additional entries for those words in the study lacking entries, and removing 70 entries for punctuation marks. We will hereafter refer to this augmented resource as "CMU-A".

Pseudowords

Pseudoword design, recording, and preparation are far more labor-intensive than for words. The speaker needs to read a phonetic notation, and often requires multiple repetitions and corrections, which consumes significantly more time. In order to optimize the time spent in recording and preparation, it was decided to record more words, rather than pseudowords (still more than four times as many pseudowords as BALDEY, the next largest auditory lexical decision database with real speech).

Pseudowords for the project were generated using the software package Wuggy (Keuleers & Brysbaert, 2010), which was kindly adapted by its author to utilize the CMU Pronouncing Dictionary (Weide, 2005) to create a phonotactic (rather than orthotactic) database. This change allowed for the creation of phonotactically licit pseudowords, which is necessary when trying to create words that a native speaker is comfortable producing. The word IPA transcriptions were input into Wuggy to ensure that phoneme and syllable makeup and transitional probabilities were comparable to the words; 11,400 of the resulting pseudowords were chosen at random as the recording list for use in MALD. The settings in Wuggy were set so that one-third of the subsyllabic constituents of the input word were swapped for other phonotactically licit segments with similar transitional probabilities. This resulted in the production of English-like "accidental gap" pseudowords, e.g., /fɪlkəm/, /gɹɑ͡ɪri/, /nuɹ/. Using Wuggy in this way, many of the resulting pseudowords also exhibit a degree of apparent morphological complexity, e.g., /tədɹe͡ɪʃɪn/, /mɑɹmsəld/, /upɪŋ/.

Speaker, recording procedure

One 28-year-old Western Canadian male phonetics undergraduate student was recorded over approximately 60 h (generally 2 h per recording session, but occasionally longer, always on weekday afternoons) in 2011 and 2012 in the Alberta Phonetics Laboratory at the University of Alberta. To help decrease effects of boredom and fatigue, the speaker took as many breaks during the recording session as he felt necessary. The speaker performed the recordings at the same time each day and if the speaker was sick recordings were not performed that day.

All word and pseudoword recordings were made with a single recording setup, which consisted of the same Countryman E6 headset microphone with 0-dB flat cap, powered by an Alesis MultiMix8 mixer/amplifier, and fed to a Korg MR-2000S studio recorder. Digital recordings were made with a sampling rate of 44.1 kHz and a bit rate of 16. The speaker was seated within a WhisperRoom sound isolation booth and read stimuli from a computer monitor positioned outside of the booth on the other side of a window. He was instructed to read words "as naturally as possible", and to read pseudowords as though they were words. Words and pseudowords were presented one at a time using E-Prime (Schneider et al., 2012). Words appeared in their standard spelling as listed in CMU-A. Pseudowords appeared in an IPA phonemic transcription with stress indicated. Both word and pseudoword lists were randomized and then sectioned into individual lists for the recording sessions.

The speaker recorded all words over 15 sessions, followed by the pseudowords in ten sessions. Words were produced once each. Pseudowords were produced with at least three repetitions (but taking care to avoid the sing-song list intonation that often accompanies items produced multiple times), so that the most fluent word-like production

could be selected. An experimenter monitored the recording for errors or disfluencies, noting problem items to be re-recorded after completion of all the item sessions.

From the base word list of 28,510 words, 1709 words or 5.99% of the full list were excluded due to mispronunciations, false starts, or disflunecies leaving 26,801 total words. Once all items were recorded and verified, the Penn Forced Aligner (Yuan & Liberman, 2008) was used to create a rough word-level alignment of each item, and these were hand-corrected. The items were then sectioned into individual sound files at zero-crossings, and each was normalized to 70 dB mean intensity using Praat's `Scale intensity` function.

Sixty-seven sets of 400 words each were matched with 24 sets ("a" - "x") of 400 pseudowords in a rotating fashion to ensure that each word set appeared with at least two different pseudoword sets. This resulted in a total of 134 randomized experimental lists each containing 800 items.

Item markup

Both word and pseudoword phone level transcriptions were time-aligned with the audio recordings once again using the Penn Forced Aligner (Yuan & Liberman, 2008). For words, a custom pronunciation dictionary was used, tailored to include all words in the study, and including a number of common pronunciation variants as well (where in each case the closest variant to the acoustics of the recording was selected by the aligner). For pseudowords, only the specific pronunciation for that pseudoword was given to the aligner for each item, except that multiple stress options were provided for each vowel such that the aligner selects the most appropriate stress assignments according to the fit of the acoustics to its models. The aligner selected at most one stressed vowel, and at most one secondarily stressed vowel. Phone boundaries for both words and pseudowords were then hand-corrected by trained phonetics research assistants. The resulting TextGrid files (Praat (Boersma & Weenink, 2011) text annotation files) are provided with each stimulus recording, indicating the approximate start and end points of each phone. Example audio files and TextGrids are illustrated in Fig. 1.

Listeners

In this section, we describe the general listener characteristics for the MALD data set. One response each for the entirety of the items requires 67 experimental runs. For this first phase of the MALD data, we collected a minimum of four responses per item, which requires a minimum of 268 experimental sessions. While having only four responses per item results in a data set that has low item power, this criterion was selected to give us a base of responses for general analysis of the lexicon. Item power will grow as further collection is completed and additional data is added to the dataset.

The listeners for MALD were 231 monolingual native Canadian English speakers who completed 284 total sessions. Listeners were recruited from the University of Alberta, Department of Linguistics participant pool and

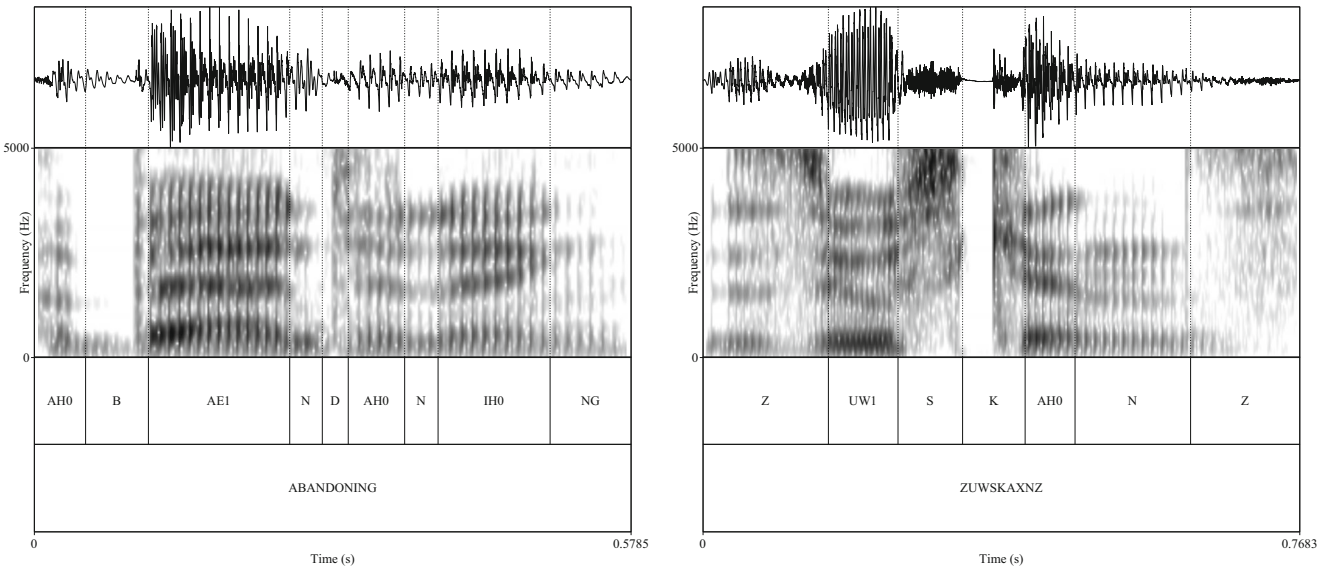

Fig. 1 Sample alignments of one word and one pseudoword from the MALD stimuli. *From top*, waveform, spectrogram, original aligned phone boundaries, item boundaries

received course credit for their participation. Listeners were permitted to participate up to three times, but never received the same words or pseudowords.

The listeners were 180 females, 51 males, aged 17–29 (mean 20.11, SD 2.39). Listeners 30 and over were not included in the present data set but will be included in a later data set. Prior to the experiment, listeners completed a demographic questionnaire, and received a rudimentary audiometric evaluation (described below). The questionnaire was used to sort participants into the relevant databases (e.g., native monolingual listeners went into the database described here), and the hearing evaluation provides further listener data for distribution with the database; it was not used in any way to exclude potential listeners. Listener characteristics are described in the analysis section below.

Experimental procedure

Each experimental session had three parts. First, a hearing evaluation was conducted. We used a Maico MA25 audiometer to present a 20-dB SPL pure tone to each ear at 500, 1000, 2000, and 4000 Hz. The participant faced away from the experimenter and equipment, and raised a hand to indicate when they had heard the tone. We recorded whether or not each tone was detected in each ear in the subject's background information.

Second, the participants began the automated auditory lexical decision task in a noise-attenuating sound booth. This was implemented using E-Prime 2.0 Professional (Schneider et al., 2012) with a Serial Response Box, both from Psychology Software Tools, Inc. Listeners were presented with 400 words and 400 pseudowords (with very slight qualifications to follow below) for each experimental session. Lexical decision sessions lasted between 20 and 25 min. Stimuli were presented over MB Quartz QP805 headphones calibrated with a 1-kHz tone to a level of 81 dB (±1 dB). This level was intended to be loud but comfortable and safe, roughly the level of a telephone dial tone, enabling comparison with a future MALD release currently being collected, which includes older listeners with various degrees of hearing loss. The calibration of the headphones was performed using a sound level meter (EXTECH 407750) with a 2cc-coupler, which simulates the resonant frequencies of the ear canal.

Each experimental session at the computer began with a set of background questions, responded to by the participant. Listeners were then provided with brief instructions where they were asked to decide whether a given item was a word of English, and to press a button with their dominant hand to select "word", and with their non-dominant hand to select "not a word". No feedback was provided during the experiment. Each trial was preceded by a 500-ms "+" fixation mark at screen center, and participants were given 3000 ms from stimulus onset to respond (mean stimulus duration = 582.15 ms; SD = 136 ms). Third, following completion of the experimental block, each listener provided additional demographic and language background information orally to the experimenter.

The stimulus presentation software E-Prime (used here to provide millisecond accuracy in response timing (Schneider et al., 2012)) operates only under Windows operating systems. The case-insensitivity of the Windows environment resulted in sound-file naming conflicts for eight-word/pseudoword pairs whose WAV files were spelled with the same letters (words "brown", "flaws", "flows", "gray", "owl", "pays", "says", "shawl"; and pseudowords [bɹo͡ʊn], [fla͡ʊs], [flo͡ʊs], [gɹa͡ɪ], [o͡ʊl], [pa͡ɪs], [sa͡ɪs], [ʃa͡ʊl], respectively). Although both members of each pair were intended to be presented in their corresponding word and pseudoword lists, only the word member was presented in both cases. This has resulted in these eight words being represented more frequently than the other words in the data set, and these eight pseudowords were not presented at all. For this reason, a small number of experimental runs contained slightly more than 400 words, and slightly fewer than 400 pseudowords.

Dataframe

A broad range of questions can be immediately addressed with the data provided in MALD, including questions germane to language processing lexical access, auditory perception, acoustic cue weighting (production and perception), individual variation, and the interactions of all of these. In what follows, we present a description of the information available in the final data files and some summary properties of the listeners, items, and responses. We then present two sample analyses investigating effects of frequency and modality on word recognition.

The MALD data frame contains listener data, item data, and the collected response data. Here we provide descriptions of all the compiled variables of each data type. Where appropriate, we also provide means and standard deviations. Variables marked with an asterisk (*) are included mainly for comparison with other data sets such as extensions of MALD which are currently being collected. These variables may be homogeneous throughout this data set, e.g., `NativeLang`, which has only the value "English" for this MALD release.

Listeners

This data set contains summary information about each participant. In addition to the data included in MALD, nine

of 256 experimental runs (from eight unique listeners) were excluded due to poor accuracy (<60% for all items). The list below provides descriptions of the individual variables. Table 1 provides a summary of the numeric variables and their distributions.

`Subject`: The listener's identifier. These are not ordered or consecutive.

`NumSess`: The number of experimental sessions the listener participated in. Listeners were permitted to participate up to three times. Eight (3%) subjects participated three times; 37 (16%) subjects participated twice; and 186 (81%) subjects participated only once.

`Age`: The listener's age at first participation. The mean age for subjects was 20.11 and the standard deviation was 2.39 years.

`Sex`: The listener's self-reported sex. Females: 180 (78%); males: 51 (22%).

`Handedness`: The listener's dominant hand. This was also the hand that the listener used to indicate words. Left-handed: 26 (11%); right-handed: 205 (89%).

`HearingScore`: a four-digit indicator of the listener's hearing evaluation results, with each digit indicating detection results at 20 dB SPL at 500 Hz, 1 kHz, 2 kHz, and 4 kHz, respectively. "0": neither ear was able to detect the tone; "1": detected the tone only in the left ear; "2": detected the tone only in the right ear; "3": detected the tone from both ears. For example, "3333" indicates the listener was able to detect tones of all four frequencies in both ears; "3032" indicates that the listener did not detect the 1-kHz tone in either ear, and detected the 4-kHz tone only in the right ear. Among our listeners, 216 (94%) detected all four tones from both ears; 15 (6%) had some degree of hearing loss evident from our screening.

Table 1 Min, max, mean, and standard deviations of some listener variables

Measure	Min	Max	Mean	SD
Background				
`Age`	17	29	20.11	2.39
`EducationLevel`	1	8	2.26	1.39
`YearsInCan`	11	29	19.96	2.46
Performance				
`Hits`	0.67	0.99	0.9	0.05
`FalseAlarms`	0.02	0.68	0.16	0.13
`Dprime`	0.86	3.35	2.47	0.47
`Beta`	0.002	9.73	0.86	1.37
`ACCRate`	0.66	0.9525	0.87	0.06
`MeanRT` (ms)	791.98	1352.68	1015.68	111.01
`WordMeanRT` (ms)	708.73	1251.7	946.22	91.47
`PwordMeanRT` (ms)	806.18	1576.17	1097.82	151.56

`EnglishProficiency*`: The listener's English proficiency as rated by the experimenters impressions of the listeners' English abilities. "1": beginner – "5": native. This variable is included mostly for comparison with nonnative listener data in the future, as all listeners were native English monolinguals.

`EducationLevel`: The listener's number of years history attending university. The mean education level was 2.26 and the standard deviation was 1.39.

`NativeLang*`: The listener's self-reported native language or languages.

Table 2 Min(imum), max(imum), mean, and standard deviations (SD) of selected item variables

Measure	Min	Max	Mean	SD
`Duration`(ms)				
Words	186	1347	578.81	136.95
Pseudowords	160	1224	591.49	132.87
`NumSylls`				
Words	1	9	2.52	1.11
Pseudowords	1	7	2.54	1.03
`NumPhones`				
Words	1	17	6.7	2.34
Pseudowords	1	17	6.93	2.23
`NumMorphs`				
Words	1	6	1.76	0.73
`PhonUP`				
Words	2	18	6.55	2.05
Pseudowords	2	10	4.66	0.92
`OrthUP`				
Words	2	19	7.67	2.23
`PhonND`				
Words	0	240	12.96	25.92
Pseudowords	0	207	5.86	14.63
`OrthND`				
Words	0	77	5.05	8.55
`PhonLev`				
Words	5.45	15.08	7.02	1.27
Pseudowords	5.56	15.01	7.1	1.26
`OrthLev`				
Words	6.15	15.17	7.78	1.24
`FreqSUBTLEX`				
Words	0	2,134,713	1612	27,394
`FreqCOCA`				
Words	0	13,780,893	8271	132,339
`FreqCOCAspok`				
Words	0	2,694,449	1890	30,508
`FreqGoogle`				
Words	0	2.3e+10	1.8e+07	2.3e+08

`OtherLangs`: Any languages other than English that the listener speaks or has studied.

`OtherLangProficiencies`: Self-ratings of proficiency in each of the listener's OtherLangs. "1": beginner – "3": advanced.

`LangAtHome*`: The languages used around the home during the listener's childhood.

`GrewUpWestCan*`: Whether the listener grew up in Western Canada.

`YearsInCan*`: The number of years the listener has lived in Canada. 13 listeners (6%) have a somewhat smaller number of years in Canada than their age due to having lived abroad for some period.

`CountryRegion*`: The listener's country of residence.

`Hits`: The proportion of word stimuli the listener correctly identified as words. The mean was 0.9, and the standard deviation was 0.05.

`FalseAlarms`: The proportion of pseudoword stimuli the listener incorrectly identified as words. The mean was 0.16, and the standard deviation was 0.13.

`Dprime`: The Signal Detection Theoretic (SDT) d′ score (Pastore & Scheirer, 1974; Abdi, 2007) for the listener. This is a measure of how well the listener discriminated words from pseudowords, measured in standard deviations, and taking into account any bias. The mean was 2.47, and the standard deviation was 0.47.

`Beta`: The SDT bias toward "word" (positive values) or "pseudoword" (negative values) responses the listener

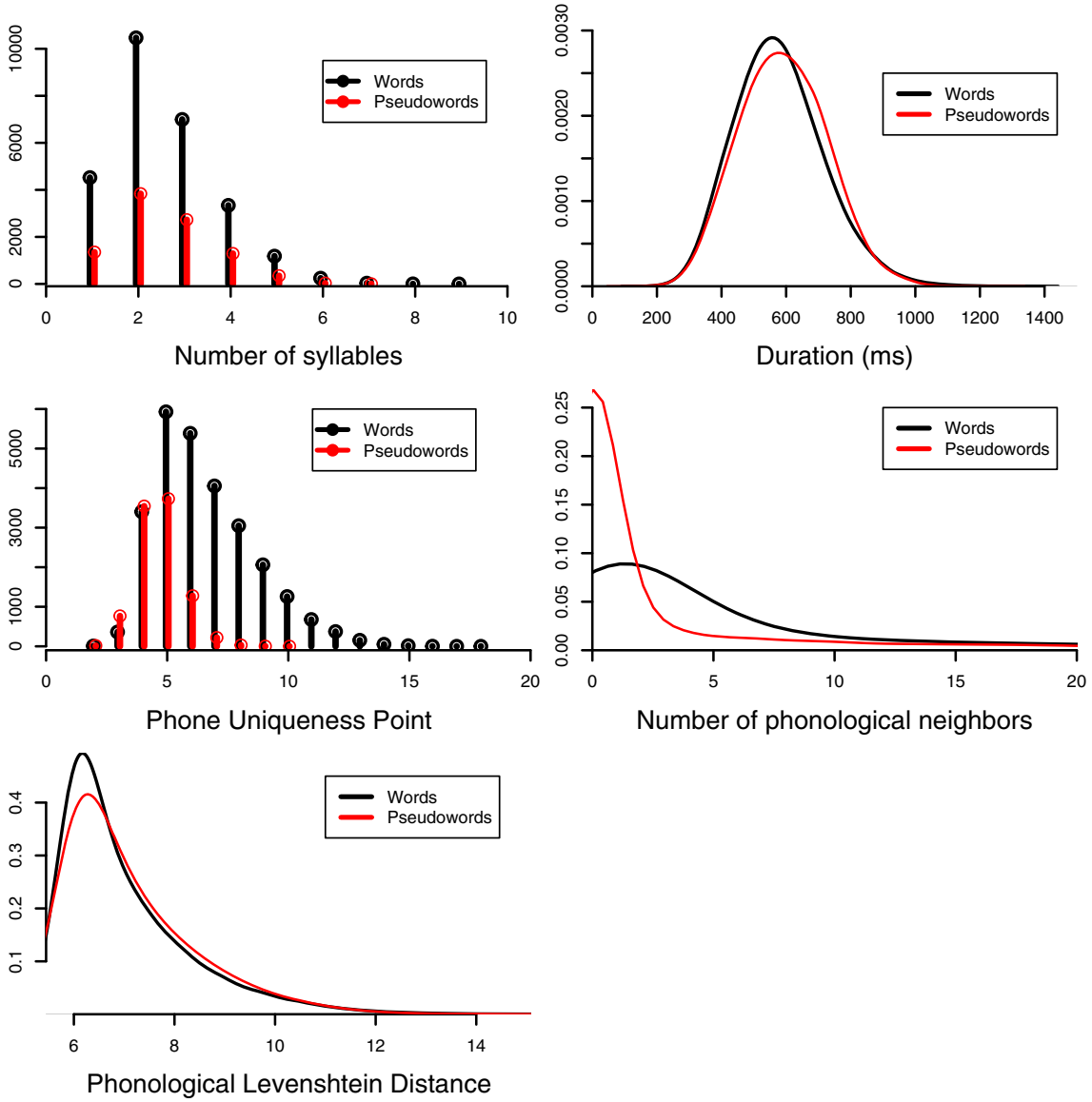

Fig. 2 Density plots for item syllable counts, durations, phonological uniqueness points, phone level neighborhood densities, and phone level mean Levenshtein distances for words and pseudowords

displayed. The mean was 0.86, and the standard deviation was 1.37.

`ACCRate`: The accuracy rate over all items (words and pseudowords).

`MeanRT`: The listener's mean response time in milliseconds.

`WordMeanRT`: The listener's mean response time in milliseconds for words.

`PwordMeanRT`: The listener's mean response time in milliseconds for pseudowords.

Items

Below we describe the item-level information bundled with the MALD release. Fields marked with a dagger (†) apply only to word items. A summary table of some of the numeric features can be found in Table 2. Figure 2 compares some of the similarities and differences between properties of words and pseudowords for several of the item variables.

`Item`: The word or pseudoword identifier.

`WAV`: The name of the sound file.

`Pronunciation`: Transcription of the item phones in the Arpabet transcription scheme.

`IsWord`: Whether the item is a word or pseudo-word.

`StressPattern`: The stress pattern of the word.

`StressCat`: The stress category of word items. The category indicates which syllable the primary stress of the word is associated with. "Medial" indicates the word has primary stress in the interior syllables of the word rather than the initial or final syllable; "multiple" indicates the word has multiple primary stresses; "secondary only" indicates the word has only secondary stress identified by CMU-A; and "none" items have all syllables listed as unstressed. The different stress categories and the number of items for each category can be found in Table 3.

`NumSylls`: The number of syllables in the item.

`NumPhones`: The number of phones in the item.

`Duration`: The duration of the item in milliseconds. The average duration of all items is 582.15 ms and the standard deviation is 136 ms. The range of durations is 160 ms to 1347 ms. As might be expected, words were produced with a slightly shorter duration (−12.68 ms) on average than pseudowords. The speaker is a trained phonetician, however, and the word/pseudoword effect on durations is small ($R^2_{(w,pw)} = 0.0017$).

Table 3 Stress patterns for words

Pattern	Frequency
Initial	17908
Medial	6580
Final	2089
Multiple	192
SecondaryOnly	5
None	14

`PhonUP`: The phone index of the phonological uniqueness point of the item within the CMU-A dictionary. The index tells which sound within the word distinguishes it from all other words. An index one greater than the number of phones is assigned if even the final sound of the item does not make it unique, e.g. 'abandon' /əbændən/ has 7 phones, but even the final phone does not yet distinguish it from 'abandoned' /əbændənd/, so the phonological uniqueness point of 'abandon' is given as 8. The mean uniqueness point phone position is 6.6 (SD: 2.1) for words and 4.7 (SD: 0.92) for pseudowords.

`OrthUP`†: The letter index of the orthographic uniqueness point of the item within the CMU-A dictionary. This is computed as for PhonUP.

`PhonND`: The number of phonological neighbors (defined as one phone edit away) for the item within the CMU-A.

`OrthND`†: The number of orthographic neighbors (one glyph edit away) for the item within the CMU-A.

`PhonLev`: Mean phone-level Levenshtein distance (Levenshtein, 1966) of the item from all entries in CMU-A. The orthographic version of this metric is shown in Yarkoni et al. (2008) to have several advantages over the traditional neighborhood density scores in accounting for lexical competition effects. We compute it here also for phones.

`OrthLev`†: Mean orthographic Levenshtein distance of the word items from all entries in CMU-A.

`POS`†: The frequency-dominant part-of-speech of the orthographic form of words according to SUBTLEX-US (Brysbaert et al., 2012). A summary of the distribution of the POS tags in the corpus are provided in Table 4.

`AllPOS`†: All parts-of-speech of the orthographic word form, in order of decreasing frequency, as given in SUBTLEX-US.

`NumMorphs`†: The number of morphemes as parsed by the PC-Kimmo two-level morphological parser (Antworth, 1995) and the Englex English morpheme sets (Antworth, 1994). It should be noted, however, that the 1200 noun–noun compounds incorporated in the item set are mostly analyzed as single morphemes by this parser. A summary of the distribution of the number of morphemes can be found in Table 5.

`FreqSubtLex`†: The frequency of the orthographic word form within the SUBTLEX-US corpus (Brysbaert et al., 2012), summing all parts of speech.

`FreqCOCA`†: The frequency of the word form within the COCA corpus (Davies, 2009).

Table 4 Summary properties of different parts of speech in MALD. Response times are for correct responses only

POS	ItemCount	ItemPerc	meanACC	sdACC	meanRT	sdRT
Adjective	4031	15.0%	0.900	0.300	955	299
Adverb	994	3.7%	0.928	0.258	967	305
Function	174	0.6%	0.892	0.310	898	267
Interjection	29	0.1%	0.865	0.343	956	355
Name	384	1.4%	0.811	0.392	986	351
Noun	13245	49.4%	0.909	0.288	940	296
Number	65	0.2%	0.936	0.244	955	305
Verb	6143	22.9%	0.928	0.259	928	287
NA	1728	6.4%	0.756	0.430	1096	375

`FreqCOCAspok`†: The frequency of the word form within the spoken language subset of the COCA corpus (Davies, 2009).

`FreqGoogle`†: The frequency of the word form within the Google Unigram corpus (Michel et al., 2011).

Figure 2 illustrates that for some of the variables, like `Syllable`, the words and pseudowords are very similar in their distributions. However, for variables like `Duration` the pseudowords are slightly, but significantly longer than the words. This is inevitable for naturally spoken pseudowords in a study of this scale. For `Uniqueness Point` the pseudowords have an earlier uniqueness point than words, which may be due to the lack of morphologically related competitors in the dictionary for pseudowords.

Another way to compare the words is to investigate the distribution of individual phonemes across the sets of words. Table 6 indicates individual phones and the count of the occurrence and percentage of the total of those phones in both the word and pseudoword lists. It can be seen that the distribution of phones is largely similar across both lists. The process Wuggy uses appears to reliably maintain the relative phone frequencies. One difference of note is that /ɚ/ does not occur in the pseudowords as a transcribed phoneme. It does occur, however, in the pseudowords as the sequence /əɹ/, which would help account for the higher percentage of /ə/ and /ɹ/ in the pseudowords. We also note that there is no occurrence of /ɔ/ in the pseudowords. The /ɔ/-/ɑ/ merger is a feature of the Western Canadian English dialect, and our speaker does not produce the distinction. However, the CMU-A transcriptions do distinguish the two sounds, thus the words reflect this vowel in their transcriptions.

Table 5 Number of morphemes in word items

Number of morphemes	1	2	3	4	5	6
Number of words	9776	11,571	2912	408	38	1

Responses

The final data set contains the trial-by-trial responses merged with the Listener and Items data sets. The list below only contains the columns not already described in the previous two data files, with `Item` and `Subject` repeated as the identifier variables across the three data sets.

`Experiment*`: "MALD1_sR" is the identifier for this release of MALD.

`Trial`: The trial number within the experimental session.

`Session`: The experimental session of the item.

`List`: The list identifier (the word list or pseudoword list) from which the item came. Each word list appeared in two different sessions in order to pair the words with two different pseudoword lists.

`WordRunLength`: The number of consecutive words or pseudowords in a row at the point of this trial in the experiment. For example, "1" indicates there was a word/pseudoword switch at this item, and "3" indicates this is the 3rd word or pseudoword in a row. Word/pseudoword item selection was randomized, making long runs of words or pseudowords possible. The average word run length was 1.99 and the standard deviation was 1.4. The run lengths followed the expected geometric distribution.

`ExperimentRunID`: A unique identifier for this particular experimental run (subject + session combination).

`RT`: Response time in milliseconds measured from item onset. The average response time is 1016.96 ms and the standard deviation is 345.02 ms.

`ACC`: Accuracy of the response where TRUE is a correct response and FALSE is an incorrect response. The average response accuracy is 87.37% and the standard deviation is 33.22%.

Table 6 Total phone token counts and percentages, in words and in pseudowords

IPA	Arpabet	Word	Pword
ɑ	AA	3183 (1.77%)	1366 (2.05%)
æ	AE	4220 (2.35%)	1393 (2.09%)
ə	AH	17268 (9.61%)	8065 (12.13%)
ɔ	AO	2034 (1.13%)	0 (0.00%)
a͡ʊ	AW	784 (0.44%)	448 (0.67%)
a͡ɪ	AY	2716 (1.51%)	1308 (1.97%)
b	B	3673 (2.05%)	1847 (2.78%)
t͡ʃ	CH	1062 (0.59%)	654 (0.98%)
d	D	7791 (4.34%)	2386 (3.59%)
ð	DH	144 (0.08%)	97 (0.15%)
ɛ	EH	5187 (2.89%)	2615 (3.93%)
ɚ	ER	5491 (3.06%)	0 (0.00%)
e͡ɪ	EY	3599 (2.00%)	1023 (1.54%)
f	F	3112 (1.73%)	991 (1.49%)
g	G	2050 (1.14%)	967 (1.45%)
h	HH	1326 (0.74%)	589 (0.89%)
ɪ	IH	11777 (6.56%)	4357 (6.55%)
i	IY	6171 (3.44%)	2157 (3.24%)
d͡ʒ	JH	1387 (0.77%)	378 (0.57%)
k	K	8416 (4.69%)	3515 (5.29%)
l	L	9802 (5.46%)	3336 (5.02%)
m	M	5559 (3.10%)	2017 (3.03%)
n	N	12013 (6.69%)	4624 (6.95%)
ŋ	NG	2817 (1.57%)	1114 (1.68%)
o͡ʊ	OW	2451 (1.36%)	845 (1.27%)
ɔ͡ɪ	OY	268 (0.15%)	14 (0.02%)
p	P	5655 (3.15%)	1713 (2.58%)
ɹ	R	9593 (5.34%)	4512 (6.78%)
s	S	11484 (6.39%)	4614 (6.94%)
ʃ	SH	2373 (1.32%)	913 (1.37%)
t	T	12532 (6.98%)	3964 (5.96%)
θ	TH	616 (0.34%)	124 (0.19%)
ʊ	UH	481 (0.27%)	125 (0.19%)
u	UW	1842 (1.03%)	618 (0.93%)
v	V	2363 (1.32%)	923 (1.39%)
w	W	1621 (0.90%)	695 (1.05%)
j	Y	1068 (0.59%)	279 (0.42%)
z	Z	5511 (3.07%)	1874 (2.82%)
ʒ	ZH	165 (0.09%)	43 (0.06%)

Response summary

The data set contains 227,179 total responses, from all listeners, including words and pseudowords. Of these responses, participants were on average 87.37% accurate with a standard deviation of 33.22%. When words are considered separately, the mean accuracy is 90.12% with 29.84% for the standard deviation. The mean accuracy for pseudowords was 84.62% and the standard deviation 36.07%. As expected, these averages indicate that participants were more accurate for the words than for the pseudowords. Participant mean response latencies were 1017 ms with a standard deviation of 345.02 ms. When words are considered separately, the mean response latency is 950.91 ms with 303.9 ms for the standard deviation. The mean accuracy for pseudowords was 1083.12 ms and the standard deviation 370 ms. As expected, these averages indicate that participants were faster to respond to words than to pseudowords.

Figure 3 provides a brief summary of the accuracy and response latency responses. The figure on the left illustrates individual participants plotted for the accuracy to the words against their pseudoword accuracy. This plot indicates that there is a general cluster of participants in the upper right, where we would expect participants that were generally accurate for both the words and pseudowords. This plot also illustrates that there were participants who had relatively high accuracy for words or pseudowords and relatively low accuracy for the other. The second plot in this figure illustrates the distribution of response latencies for both words and pseudowords. Here, as would be expected from the body of literature, we see that words are responded to faster than pseudowords, and that the deviation from the mean is also smaller.

Example analyses

Large datasets such as MALD enable certain analyses which are out of scope of smaller, targeted experiments. In the remainder of this section, we present the results of four such analyses. First, we look at frequency, a major factor in lexical decision tasks, and compare frequency estimates from various sources in their ability to predict participant accuracy and response latencies. Second, we draw from two megastudies, MALD and English Lexicon Project, to investigate differences in effects various predictors have in the auditory versus the visual modality. Third, we briefly describe one published study which makes use of the MALD dataset (Schmidtke et al., 2018) to investigate the processing of compound words. Finally, we briefly summarize the results of a replication of Goh et al. (2016) using the MALD stimuli.

Corpus frequency comparison

In our first example, we examine the effects of frequency in the MALD data and investigate how predictive individual corpora are for the MALD data. Frequency is one of the most robust effects in psycholinguistic research. In

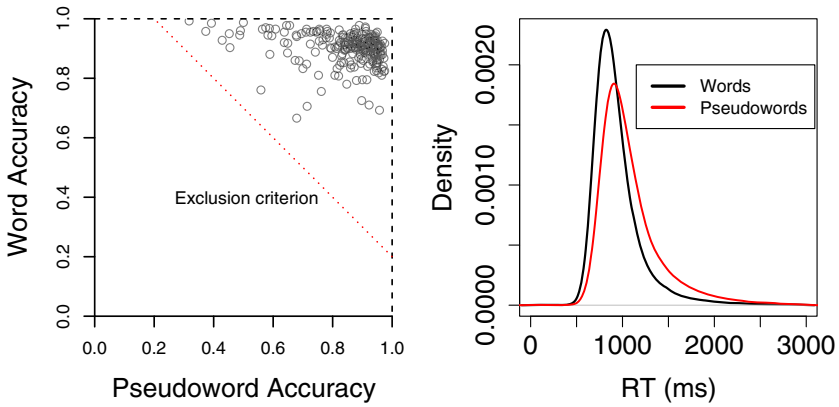

Fig. 3 Left: accuracy rates by listener, for words and pseudowords (listeners with total accuracy below 60% were excluded); Right: response latency density functions of words and pseudowords

these studies, researchers employ some corpus count to approximate the relative ambient exposure of selected items, and this predicts a good portion of the variance in how fast listeners respond to those items. The nature of this frequency effect, however, is not well understood. We compared frequency counts compiled from three large US English corpora: the `Google` Unigram corpus (Michel et al., 2011), the Corpus of Contemporary American English (Davies, 2009); which was divided into two corpora: the full corpus (`COCA`), and the subset containing only spoken materials (`COCAspok`), and a movie subtitle corpus of American English (Brysbaert et al., 2012, SUBTLEX-US).

We used linear mixed-effects modeling to compare the frequency effects from the various corpora on modeling the MALD data (Baayen et al., 2008). We created baseline models of the response latencies for correct responses and the accuracy for words. In these models, we did not explore any potential interactions. We included the following control predictors: `PhonND`, `PhonUP`, `Duration`, `Trial` and `WordRunLength`. All of the predictors were scaled (with their center at 0) and all but `Trial` and `WordRunLength` were logged for normality in their distributions. We added 1 to each frequency value and PhonND to ensure nonzero values for the logarithmic transform. We included random effects for `Subject` and `Item` with random slopes for `PhonND`, `Duration`, `Trial` and `WordRunLength` by `Subject` in the response latency model. Random slopes that did not significantly improve the model fit were not included in the model. We tested `Age` and `Sex` as control variables in the baseline model but they were not statistically significant predictors and were excluded in subsequent models. In the accuracy model, partly due to the fact that there were so few responses per item and to help address convergence issues, we did not include `Item` as a random effect, or random slopes. We then created four separate models, each expanding the baseline model by including one of the frequency measures. Models with the lowest Akaike Information Criterion (AIC; Akaike (1973)) are the models providing the strongest fit to the data. Table 7 consists of the AICs for each corpus or corpus subset.

The results of the frequency corpus comparison indicate different results depending on the dependent variable analyzed. The models were statistically compared to one another using ANOVA. All within variable comparisons were significant except for the comparison of `SubtLex` and `COCA` in the response latencies. The response latency comparisons indicate that the AIC for the subtitle-based corpus, `SubtLex`, and `COCA` provided the best fit to

Table 7 AIC goodness-of-fit ratings for the word frequency models yielded by counts from various corpora

Model	AIC_RT	AIC_ACC
Baseline	−13956.00	66843.82
Google	−16280.46	61692.71
COCA	−16479.90	61240.61
COCAspok	−16390.59	61120.17
SubtLex	−16477.47	60548.80

Table 8 Results of the model using COCA frequency

	Estimate	Std. error	*t* value
Intercept	6.894	0.007	1042.111
Duration	0.073	0.001	54.202
COCA	−0.074	0.002	−34.622
PhonND	0.015	0.001	11.399
Trial	−0.022	0.002	−10.877
PhonUP	−0.012	0.001	−12.219
WordRunLength	−0.011	0.001	−10.584

Note all numeric predictors have been scaled

the data. The results of the linear mixed-effects model for response latencies using the COCA frequency-based measure can be found in Table 8. The other models are essentially identical to this one. In the accuracy models, we find that COCAspok is more predictive than COCA, though SubtLex provides the best overall fit. In all cases Google provided the worst fit to the data.

By way of further comparison, we have also provided a correlation matrix of the four frequency measures in Table 9. This table indicates that there is a strong relationship between SubtLex and COCAspok, which squares with the similar content of the two corpora. COCA and COCAspok are understandably strongly correlated as COCAspok is a subset of COCA. Interestingly, the correlation between COCA and SubtLex is weaker than most of the other correlations even though they equally account for the latency data. We suspect that this may be due to the fact that each corpus captures different genres of language better. We also suspect that the reason both subtitle corpora (COCAspok and SubtLex) perform better in the accuracy models is that they better represent the lexicon that most participants are familiar with and use in their day-to-day interactions.

Comparison of auditory and visual lexical decisions

Although both visual and auditory word processing literatures are longstanding and developed (e.g., visual: Forster (1976), Coltheart et al. (1977), Forster et al. (2003), and Balota et al. (2006); auditory: (Jusezyk & Luce, 2002; Smits et al., 2003; Cutler, 2012)), there is surprisingly little direct comparison of the two modalities. There may be a great deal of shared architecture in the systems involved in both types of recognition (Bradley & Forster, 1987), but (Taft, 1986), (Goh et al., 2016) and (Ferrand et al., 2017) are the only behavioral studies we were able to uncover which compare the two modalities directly (though see Chee et al. (1999) and Rayner and Clifton (2009), for neuroimaging studies). Large visual databases have existed for some time, and now with the addition of MALD, these auditory/visual comparisons are quite accessible.

In the comparison that follows, we explore how the patterns of duration, word frequency, neighborhood density, uniqueness point, trial number, and word run length might differ between the modalities. In addition, we compare whether phonological neighborhood density (as computed with phone-level representations of the words and the dictionary entries) differs from orthographic neighborhood density in their fit to the data ($\text{corr}_{phonND,orthND} = 0.73$). Since the two neighborhoods are computed in dimensions that are tailored to the two styles, a portion of the competition effects might be better represented in the respective units. On the other hand, since competition effects are presumed to be a property of the organization of a single lexicon employed in both reading and listening, it might be expected that one or the other of the ND measures would better model this property of the lexicon and produce better model fits for both modalities.

Table 9 Correlation matrix of the four frequency variables calculated from the different corpora

	Google	COCA	COCAspok	SubtLex
Google	1			
COCA	0.974	1		
COCAspok	0.926	0.967	1	
SubtLex	0.695	0.75	0.861	1

We subsetted the MALD and the visual lexical decision portion of the English Lexicon Project (Balota et al. 2007, henceforth "ELP") data sets to include only word items common to both corpora (25,389 unique words, including 95% of MALD words, and 63% of ELP words). Before merging the two data sets, we computed word run length (recall that this is the number of words or pseudowords that occurred in a row at the point of the item) for ELP, and in both corpora we excluded responses faster than 200 ms or slower than 3000 ms, and included only correct responses for our response latency analysis. The resulting data set contained 97,069 MALD responses from 231 subjects and 750,299 ELP responses from 815 subjects. Apart from the size differences, it should be noted that although we are comparing only the words the two corpora had in common, the pseudoword sets used in the two experiments were different.

We performed subtractive nested AIC goodness-of-fit tests for linear mixed effects models of response time, comparing models using phonological neighborhood density with those using orthographic neighborhood density, and comparing global models including both corpora with individual models for each corpus. The models are aimed at comparing the two perception modalities and the two measures of lexical competition.

All variables in the models were natural log transformed to more normally distribute their residuals with respect to the response times, and then converted to z-scores for comparable scaling across the variables, with the exception of Trial and WordRunLength which were scaled only. As before we smoothed both PhonND and COCA frequencies by incrementing by 1, to shift the variable ranges into the domain of the log transform.

Since we are interested in the differences between the two modalities, we included interaction terms between Corpus and Duration (the duration of the stimulus recording in the auditory experiment), COCA (COCA corpus frequency - as this provided the best fit in the previous latency analysis), PhonND (phonological neighborhood

density), and PhonUP (phonological uniqueness point). Additional control variable fixed effects were also included for Trial and WordRunLength. Random intercepts were included for Item, as well as random intercepts for Subject with slopes for COCA and PhonND. Additional slopes by subject or item became computationally infeasible, and were not critical to our intended comparisons.

The planned model comparison procedure was to remove fixed interactions and then fixed effects, selecting at each point the more parsimonious model unless the model with additional parameters was significantly better fitting (by anova model comparison in R; all lmer models were fit by maximum likelihood to enable this comparison) as indicated by the AIC. In this case, the fully specified model, with all the aforementioned interactions, was by far the best fitting, several hundred AIC points lower than the next best model (lower AIC values index better model fit). Additionally, the phonological neighborhood density (PhonND) provided the best fit to the data in both the omnibus model (AIC_{phonND} = 1,911,609; AIC_{orthND} = 1,911,897) as well as the two individual corpus models (auditory data set: AIC_{phonND} = 177,040.2, AIC_{orthND} = 177,149.2 ; visual data set: AIC_{phonND} = 1,721,626, AIC_{orthND} = 1,722,601). This suggests that even in the visual word recognition modality, competition effects may be phone-based rather than glyph-based. Since the phonND models were the best fitting in all cases, we present only results from models with that variety of neighborhood density variable.

Table 10 presents the parameter estimates for the omnibus model comparing the auditory and visual lexical decision data. Separate models were also fit for each corpus, and corroborate the patterns shown here. The effect of corpus reveals that auditory lexical decision is slower on average than visual lexical decision (mean RT_{MALD} = 940 ms; mean RT_{ELP} = 759 ms). Longer stimulus duration recordings, and items with later uniqueness points also produced longer response times overall. Later trials, however, as well as longer word run lengths (with the latter effect several times larger than the former) resulted in faster responses. Higher-frequency items, and items from higher density neighborhoods also had faster responses on average. In the interactions by corpus, stimulus recording duration has a much larger effect in MALD than for ELP, which squares with the fact that MALD is the only corpus in which the recordings were actually presented. The effect of duration in ELP is mediated by the correlation of the number of letters in the word with the length of the spoken word, weakening the association by some measure. The facilitation afforded by frequency appears to be nearly twice as much in the visual domain as in the auditory domain, as is the effect of neighborhood density. Surprisingly, the effect of uniqueness point is also somewhat larger in the visual domain.

Table 10 Results of the model comparing both corpora. Note all numeric predictors have been scaled

	Estimate	Std. error	*t* value
Intercept	−0.058	0.016	−3.530
MALD	0.638	0.035	18.316
Duration	0.072	0.002	30.364
COCA	−0.242	0.003	−88.763
PhonND	−0.076	0.003	−22.334
PhonUP	0.040	0.002	18.069
Trial	−0.044	0.001	−50.408
WordRunLength	−0.070	0.001	−86.845
MALD:Duration	0.134	0.003	38.499
MALD:COCA	0.120	0.005	23.512
MALD:PhonND	0.116	0.006	18.883
MALD:PhonUP	−0.078	0.003	−24.171

Conceptual relations of compounds

Schmidtke et al. (2018) investigate the effect of competing conceptual relations in the recognition of compound words. In experiment 2, the authors explore these effects in the auditory domain using 412 English compounds from the MALD dataset and in experiment 3 they make use of an auditory compound lexical decision dataset with 426 English compounds. In the analyses of both datasets they find effects of the entropy of conceptual relations (effect sizes of 56 ms and 37 ms, respectively). They also found no effect of family size for these items. Interestingly, they find different effects for Left-Whole and Right-Whole semantic similarity. The authors attribute this difference to the fact that in the MALD dataset the compound words are mixed in with many non-compound words while the dataset in experiment 3 is exclusively compounds (words and pseudowords).

Replication of semantic richness effects

In the final example, we show how MALD recordings can be used as stimuli in creating targeted experiments and how MALD data set corresponds to yet another targeted experiment's findings. We extracted all of the items from the Goh et al. (2016) study which overlapped with the items in the MALD dataset, 442 nouns, and selected a random sample of 442 MALD pseudowords. We ran a separate experiment with an additional group of 25 English speaking participants with only these items from the MALD recordings. Our analysis was an attempt to replicate the procedure described in Goh et al. (2016), in

which hierarchical linear regression was used on reaction times that were z-scored per participants and then averaged by item. The results showed that the same hierarchical model (out of four tested) was the best fit to the data in both studies. For semantic richness variables, we gathered the same measures as those used by Goh et al. (2016): concreteness (Brysbaert et al., 2014), valence and arousal (Warriner et al., 2013), number of features (McRae et al., 2005), semantic neighborhood density (Shaoul & Westbury, 2010), and semantic diversity (Hoffman et al., 2013). Identical patterns of significance and effect direction were noted for all semantic richness variables. The effects reported were largely replicated. For lexical predictors, same patterns were observed for duration, frequency, phonological uniqueness point (non-significant), and a structural principal component (consisting of phonological neighborhood density, phonological Levenshtein distance, number of phones, and number of syllables). the only exception was number of morphemes, which was not significant in our replication, but was significant in the Goh et al. (2016) study (note however that the effects of this variable were unstable, and non-significant in the semantic categorization experiment conducted as part of the same study). For semantic richness variables, identical patterns of significance and effect direction were noted for all variables (concreteness, valence, number of features, semantic neighborhood density, and semantic diversity). We did not see a significant improvement in the model when quadratic valence was added to the model. The Table 11 in the appendix contains the coefficients from the analyses.

General discussion

In this paper, we have described the methods and characteristics of the first release of MALD and have illustrated some basic questions that can be investigated using the data set. It provides formidable subject-level and item-property-level (e.g., frequency) statistical power, and as further data is collected, item-specific power (i.e., how listeners recognize specific individual words) will be solidified as well.

As an illustration, we performed three sample analyses using the MALD data set or MALD stimuli and summarized another study which has already made use of the MALD dataset. The first analysis investigated four different sources of calculating frequency. Our findings were in line with other research, and indicated that a subtitle-based frequency count (SUBTLEX) best explains the frequency effects on response times, even when compared to a balanced corpus like COCA or the spoken subset of COCA, which both should have similarity to the subtitle corpus. This corroborates the findings from Ernestus and Cutler (2015). In the second analysis, we compared the data from MALD to the data from the ELP with several standard statistical predictors. In the comparison of auditory lexical decision to visual lexical decision we were able to replicate the general findings of neighborhood density for both visual and auditory modalities, where dense neighborhood words are recognized more rapidly. We also found that phonological neighborhood density, rather than orthographic neighborhood density, better predicts not only the auditory lexical decision results but also the visual lexical decision data. The results from both of these analyses deserve more detailed exploration. The fourth analysis successfully replicated an existing study using the MALD stimuli. We believe that they provide inviting demonstrations of some of the research that can be done with these datasets.

As with any single task methodology, there are limitations to auditory lexical decision. However, since such a large body of the literature has employed this task, we feel that this is a reasonable place to begin, with the hope that researchers will expand such endeavors to other tasks. We also recognize that there is a bias in language processing research toward visual materials, not only in the megastudies that are available but also more generally (Jarema et al., 2015). We hope that by making this data set available, more researchers who might otherwise avoid investigations of speech will expand their domains of interest, and we encourage other speech and psycholinguistic researchers to consider what other theoretically significant but underdeveloped areas of psycholinguistic theory might benefit from megastudies of their own, such as conversational speech (Tucker & Ernestus, 2016).

Availability The corpus is available publicly and can be downloaded here: http://mald.artsrn.ualberta.ca/

As previously noted, data collection for MALD is ongoing and will be continually updated on the public website as more data is available. The data sets will be associated with version numbers so that when the current version of MALD1 (v1.0) is updated with additional data the version number can be used to indicate which version of the data has been used.

In the spirit of open science, we encourage researchers who calculate other variables for the data set as part of their research to share them with us so that we can add them to future MALD versions. We also encourage researchers who design experiments using the MALD stimuli to share their data publicly.

Acknowledgements This research was funded by SSHRC Grant #435-2014-0678 and by a University of Alberta Killam Research Grant, both to the first author. It also benefited greatly from planning consultation with R. Harald Baayen, and organizational, subject-running, coding, and markup contributions by Kara Hawthorne,

Danielle Fonseca, Catherine Ford, Pearl Lorentzen, and Katelynn Pawlenchuk. Thanks also to Emmanuel Keuleers for adapting Wuggy for our pseudoword creation. Correspondence may be addressed to Benjamin V. Tucker, 4-32 Assiniboia Hall, Department of Linguistics, University of Alberta, Edmonton, Alberta, T6G2E7, Canada (e-mail: bvtucker@ualberta.ca).

Appendix

Table 11 Standardized regression coefficients for item-level hierarchical regression analyses from the replication of Goh et al. (2016)

Model 1: Lexical variables (Control)	
Word duration	0.00***
log subtitle word freq	−0.07***
Phonological UP	0.03
Structural principal component	−0.14**
Num. of morphemes	−0.02
Adjusted R^2	0.2871***
Model 2: Semantic richness variables	
Concreteness	−0.34***
Valence	−0.05**
Arousal	−0.03.
Number of features	−0.01**
Semantic neighborhood density	0.12
Semantic diversity	−0.11
ΔR^2	0.0671***
Model 3: Quadratic valence	
Valence^2	0.01
Δ R^2	−0.0015
Model 4: Valence × arousal	
Valence × arousal	−0.01
Arousal × valence^2	0.00
Δ R^2	−0.0031

References

Abdi, H. (2007). *Signal detection theory (SDT). Encyclopedia of measurement and statistics* (pp. 886–889).

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Czaki*. Budapest: Akademiai Kiado.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461. https://link.springer.com/article/10.3758/BF03214334

Antworth, E. L. (1994). Morphological parsing with a unification-based word grammar. In: *Proceedings of the North Texas Natural Language Processing Workshop* (pp. 24–32).

Antworth, E. L. (1995). User's guide to pc-kimmo version 2. Página web]. Disponible en http://www.sil.org/pckimmo/v2/doc/guide.html

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. http://www.sciencedirect.com/science/article/pii/S0749596X16302467

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM). Linguistic Data Consortium*. Philadelphia: University of Pennsylvania.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). *Handbook of psycholinguistics* (pp. 285–375). Academic Press, ch. Visual Word Recognition (Ch. 9).

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., ..., Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 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. Visual word recognition volume 1: Models and methods, orthography and phonology, 90. https://books.google.ca/books?hl=en&lr=&id=uco5lasTR2oC&oi=fnd&pg=PA90&dq=Megastudies:+What+do+millions+(or+so)+of+trials+tell+us+about+lexical+processing&ots=azU5QTJobe&sig=CFZNClqnhGkjOgPTXsYpZPTiKYc

Boersma, P., & Weenink, D. (2011). Praat, a system for doing phonetics by computer. www.praat.org

Bradley, D. C., & Forster, K. I. (1987). A reader's view of listening. *Cognition*, *25*, 103–134.

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

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Campbell, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Chee, M. W., O'Craven, K. M., Bergida, R., Rosen, B. R., & Savoy, R. L. (1999). Auditory and visual word processing studied with fMRI. *Human Brain Mapping*, *7*(1), 15–28.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In Dornic, S. (Ed.) *Attention and Performance VI* (pp. 535–555). Hillsdale: Lawrence Erlbaum Associates. http://www.maccs.mq.edu.au/max/cv/#four.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: will we able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*(1-3), 65–70. http://pubman.mpdl.mpg.de/pubman/faces/viewItemFullPage.jsp?itemId=escidoc:68678

Cutler, A. (2012). *Native listening: language experience and the recognition of spoken words*. Cambridge: MIT Press.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (19902008+): design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190.

Dufau, S., Grainger, J., & Ziegler, J. C. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 1117.

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1469–1488. https://doi.org/10.1080/17470218.2014.984730

Ferrand, L., Mot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., . . . , Grainger, J. (2017). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, 1–23. https://link.springer.com/article/10.3758/s13428-017-0943-1

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., . . . , Pallier, C. (2010). The French Lexicon Project: lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496.

Forster, K., Mohan, K., & Hector, J. (2003). *Masked priming: State of the art* (pp. 3–37). New York: Psychology Press Ch, The Mechanics of Masked Priming (Ch. 1).

Forster, K. I. (1976). Accessing the mental lexicon. In Wales, R. J., & Walker, E. (Eds.) *New approaches to language mechanisms*, (pp. 257–287). Amsterdam: A collection of psycholinguistic studies.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*, 1109–1115.

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 Categorization Megastudy. Frontiers in Psychology 7. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00976/full

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos Med*, *2*(8), 0696–0701.

Jarema, G., Libben, G., & Tucker, B. V. (2015). The integration of phonological and phonetic processing: a matter of sound judgment Jarema, G., & Libben, G. (Eds.) *Benjamins Current Topics* (Vol. 80, pp. 1–14). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/bct.80.002int. https://benjamins.com/catalog/bct.80.002int

Jusezyk, P. W., & Luce, P. A. (2002). Speech perception and spoken word recognition: past and present. *Ear and Hearing*, *23*(1), 2–40.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1457–1468. https://doi.org/10.1080/17470218.2015.1051065

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

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: a lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Language Sciences* 1, 174. http://www.frontiersin.org/language_sciences/10.3389/fpsyg.2010.00174/abstract

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*(1), 287–304.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1665–1692. https://doi.org/10.1080/17470218.2015.1022560

Kuperman, V. (2015). Virtual experiments in megastudies: a case study of language and emotion. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1693–1710. https://doi.org/10.1080/17470218.2014.989865

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics – Doklady*, *10*, 707–710.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467695/

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. http://www.sciencedirect.com/science/article/pii/S0749596X16300079

Mattys, S. L. (1997). The use of time during lexical processing and segmentation: a review. *Psychonomic Bulletin & Review*, *4*(3), 310–329. http://link.springer.com/article/10.3758/BF03210789

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M., The Google Books Team, . . . , Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

New, B. et al. (2006). Reexamining the word length effect in visual word recognition: new evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52.

Norris, D. (2013). Models of visual word recognition. *Trends in Cognitive Sciences*, *17*(10), 517–524.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. http://www.ncbi.nlm.nih.gov/pubmed/18426294

Pastore, R., & Scheirer, C. (1974). Signal detection theory: considerations for general application. *Psychological Bulletin*, *81*(12), 945–958.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology. Ohio State University (Distributor).

Radach, R., & Kennedy, A. (2013). Eye movements in reading: some theoretical context. *The Quarterly Journal of Experimental Psychology*, *66*(3), 429–452. https://doi.org/10.1080/17470218.2012.750676

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255. https://doi.org/10.1207/s1532799xssr1003_3

Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: implications for event-related potential research. *Biological Psychology*, *80*(1), 4–9.

Schmidtke, D., Gagn, C. L., Kuperman, V., Spalding, T. L., & Tucker, B. V. (2018). Conceptual relations compete during auditory and visual compound word recognition. *Language, Cognition and Neuroscience*. http://www.tandfonline.com/doi/abs/10.1080/23273798.2018.1437192

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime Reference guide*. Pittsburgh: Psychology Software Tools Inc.

Schröter, P., & Schroeder, S. (2017). The developmental lexicon project: a behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods*, 1–21. https://doi.org/10.3758/s13428-016-0851-9

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309.

Seidenberg, M. S., & Waters, G. S. (1989). Reading words aloud—a mega study. *Bulletin of the Psychonomic Society*, *27*(6), 489–489.

Shadish, W. R. (1993). Critical multiplism: a research strategy and its attendant tactics. *New Directions for Evaluation*, *1993*(60), 13–57.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393–413.

Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: a database of Dutch diphone perception. *The Journal of the Acoustical Society of America*, *113*(1), 563–574.

Taft, M. (1986). Lexical access codes in visual and auditory word recognition. *Language and Cognitive Processes*, *1*(4), 297–308.

ten Bosch, L., Boves, L., & Ernestus, M. (2013). Towards an end-to-end computational model of speech comprehension: simulating a lexical decision task. In: INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association. pp. 2822–2826. http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:1835410

ten Bosch, L., Boves, L., & Ernestus, M. (2015a). DIANA, an end-to-end computational model of human word comprehension. In: 18th International Congress of Phonetic Sciences (ICPhS 2015). University of Glasgow. http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:2181015

ten Bosch, L., Boves, L., Tucker, B., & Ernestus, M. (2015b). DIANA: towards computational modeling reaction times in lexical decision in North American English. In: Interspeech 2015: 16th Annual Conference of the International Speech Communication Association. pp. 1576–1580. http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:2230858

ten Bosch, L., Ernestus, M., & Boves, L. (2014). Comparing reaction time sequences from human participants and computational models. In: Interspeech 2014: 15th Annual Conference of the International Speech Communication Association. pp. 462–466. http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:2058455

Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W. P., Shaoul, C., & Lin, D. (2016). The Chinese Lexicon Project: a megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, 1–17. https://doi.org/10.3758/s13428-016-0810-5

Tucker, B. V., & Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, *11*(3), 375–400. http://www.jbe-platform.com/content/journals/10.1075/ml.11.3.03tuc

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329. http://www.jstor.org/stable/40063346

Warner, N., Clayton, I. D., Carnie, A., Fisher, M., Brenner, D., & Hammond, M. (2014). The effect of Gaelic initial consonant mutation on spoken word recognition. In: *Celtic linguistics conference* (Vol. 8). Edinburgh, UK, poster presentation.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Weide, R. (2005). *The Carnegie Mellon Pronouncing Dictionary* [cmudict. 0.6]. Carnegie Mellon University: http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Accessed.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53.

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: insights from the English Lexicon Project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 597.

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

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, *11*(3), 452–457. https://link.springer.com/article/10.3758/BF03196594

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics*.

Ziegler, J. C., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*(4), 779–793. http://www.sciencedirect.com/science/article/pii/S0749596X03000068