Title: **BALDEY: A database of auditory lexical decisions**

Authors:

- Mirjam Ernestus

  Centre for Language Studies, Radboud University Nijmegen & Max Planck Institute for Psycholinguistics, Wundtlaan 1, P.O. Box 310, 6500 AH Nijmegen, The Netherlands
- Anne Cutler

  The MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith South, NSW 2751, Australia, and Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands

Short title (max 40 characters): A database of auditory lexical decisions

Corresponding author:

Mirjam Ernestus

Centre for Language Studies, Radboud University Nijmegen & Max Planck Institute for Psycholinguistics, Wundtlaan 1, P.O. Box 310, 6500 AH Nijmegen, The Netherlands

Email address: m.ernestus@let.ru.nl

Telephone number: +31 24 3612970

Fax number: +31 24 3521213

Acknowledgments: We thank Harald Baayen for useful discussions on vocabulary structure and design of the database, and Tijl Grootswagers for assistance with the automatic database augmentation procedures.

**Abstract**

In an auditory lexical decision experiment, 5,541 spoken content words and pseudo-words were presented to 20 native speakers of Dutch. The words vary in phonological makeup and in number of syllables and stress pattern, and are further representative of the native Dutch vocabulary in that most are morphologically complex, comprising two stems or one stem plus derivational and inflectional suffixes, with inflections representing both regular and irregular paradigms; the pseudo-words were matched in these respects to the real words. The BALDEY data file includes response times and accuracy rates, with for each item morphological information plus phonological and acoustic information derived from automatic phonemic segmentation of the stimuli.

Two initial analyses illustrate how this data set can be used. First, we discuss several measures of the point at which a word has no further neighbors, and compare the degree to which each measure predicts our lexical decision response outcomes. Second, we investigate how well four different measures of frequency of occurrence (from written corpora, spoken corpora, subtitles and frequency ratings by 70 participants) predict the same outcomes. These analyses motivate general conclusions about the auditory lexical decision task. The (publicly available) BALDEY database lends itself to many further analyses.

Key words: Auditory lexical decision, Morphologically complex words, Frequency of occurrence, Phonological neighbors, Dutch

**Introduction**

No matter how predictable some utterances may seem to be, alternatives are always possible. Even a person who arrives every day at the same place at the same time and has on every previous occasion said "good morning" may one day begin with "quick, you must see this!", or "whatever has happened here?". And if so, listeners will respond appropriately, for the processes of spoken-word recognition operate on whatever auditory input comes in, and deliver its interpretation with a rapidity born of the massive over-learning that characterises such everyday cognitive processing.

The rapidity and the sheer ordinariness of the experience of recognising spoken words should not, however, be allowed to mask the complexity of the processing involved. Vocabularies contain, in any language, hundreds of thousands of individual stand-alone phonological word forms associated to their appropriate meanings. Crucially, these forms are not easily discriminable, because they are made up of only a handful (on average, between two and three dozen) of contrastive speech sounds. Words therefore resemble other words, and longer words contain shorter words embedded within them. The process of recognising words is one of sorting out the actually spoken input from all the other word forms for which the input also provides full or partial support.

These alternatives receive consideration by listeners, or, as word recognition researchers put it, are activated in the listener's mind. Even though word recognition proceeds so very rapidly, experiments show fleeting availability of temporarily supported words. Alternative interpretations compete with one another, and the more of them there are, the slower recognition proceeds. (See McQueen, 2007, for a succinct review of the most important issues in current spoken-word recognition research, and the most influential findings.) Because these features of the recognition task – multiple word-form activation, inter-form competition – have their origin in the composition of very large vocabularies using

very small speech-sound inventories, they are effectively universal across languages. Languages differ in whether they construct utterances uniquely of elements that can stand alone, expressing morphosyntactic relationships with stand-alone particles too (as in the languages of China), whether they make utterances only of elements that can never occur alone (as in the polysynthetic indigenous languages of Australia or North America), or whether they use a mixture of stand-alone and bound elements (as in most European languages); but in all cases, utterances will temporarily support multiple interpretations and listeners will temporarily consider them.

Spoken-word recognition researchers have established this picture largely by the use of methods specifically designed to test for the activation of alternative meanings, including (a) eye-tracking (Tanenhaus & Spivey-Knowlton, 1996), in which listeners hear spoken input while their looks, however brief, to members of a small response set of pictures or printed words are registered, (b) cross-modal priming (Zwitserlood, 1996), in which a whole spoken word or a spoken word fragment serves as prime to a test word which is identical to, related to, or a competitor of the input word, with the test word usually examined by visual lexical decision; or (c) word-spotting (McQueen, 1996), which is essentially a go/no-go form of auditory lexical decision in which listeners respond upon detecting real words embedded in nonsense strings. These are the techniques that have most firmly established that multiple interpretations of the input are temporarily available during listening, and that competition occurs in the sense that increasing support for one interpretation leads to inhibition of other interpretations (Allopenna, Magnuson & Tanenhaus, 1998, with eye-tracking; McQueen, Norris & Cutler, 1994 with word-spotting), and that the more alternative interpretations are simultaneously available, the slower recognition occurs (Norris, McQueen & Cutler, 1995 with word-spotting; Vroomen & De Gelder, 1995 with cross-modal priming).

Those multiple-activation techniques are less often employed for addressing the most fundamental questions of spoken-word recognition, such as the point at which a spoken word can be definitively recognised, or the role in recognition of lexical factors such as the word's frequency, or morphological complexity. For such questions, the spoken-word recognition researcher's favourite workhorse is – as with word recognition in the visual modality – the lexical decision task. Of course, auditory lexical decision data shows multiple activation and competition effects too; thus it is harder to reject nonwords that are still compatible with potential words than nonwords that could never be continued to become words (e.g., *shrap*, which could become *shrapnel*, versus *shrip*, which cannot be continued to become a real word; Taft, 1986). The same is true for nonwords which have been cross-spliced from real words and thus still have coarticulatory information supporting the real-word interpretation, e.g., *troot* in which the *troo-* came from an utterance of the real word *troop*, versus *troot* in which the *troo-* had originally been spoken in the nonword *trook* (Marslen-Wilson & Warren, 1994; McQueen, Norris, & Cutler, 1999). The size of the lexical neighbourhood also affects decision time for real words (Goldinger, Luce & Pisoni, 1989). All of these results suggest continuous consideration of potentially multiple possible interpretations.

Auditory lexical decision resembles visual lexical decision in showing clear effects of word frequency (Connine, Mullennix, Shernoff & Yelen, 1990), of morphological complexity (Schreuder & Baayen, 1995), and of repetition priming (Slowiaczek & Pisoni, 1986). Early uses of the auditory task are described by Goldinger (1996). These include Marslen-Wilson's (1980) experiments in support of his claims for a role in recognition of the Uniqueness Point (the point in any spoken word at which no further competing interpretations exist); again, this underlines the importance for the listener of distinguishing an incoming word from other words that it potentially could become. A consequence of this vital feature of the task is that listeners cannot (except in nonwords) issue a definitive response before the end of the spoken

input. Even if we hear *alligat-* there is no guarantee, in a lexical decision situation, that the rest of the input will be, and will only be, the syllable *-or*; it might turn out that we are hearing a nonword such as *alligatif* or *alligatoreen*. This is an important difference between the visual and auditory versions of the lexical decision task; in the visual task, participants can see at a glance whether a presented form is long or short, but in the auditory task, listeners must wait for silence to tell them that the presented form has ended. For this reason, lexical decision times for spoken real words do in fact exhibit effects of word characteristics after the uniqueness point (Goodman & Huttenlocher, 1988; Taft & Hambly, 1986), and in customary practice, the duration of each spoken lexical decision stimulus is measured, allowing response times to be calculated either from word onset or from word offset.

In both the visual and auditory versions of the task, it is of course important that the nonwords are, at least temporarily, plausible contenders as lexical items. If every nonword can immediately be rejected as a lexical candidate, participants can adopt a superficial strategy which does not require actual recognition of the real words. Implausible visual nonwords (e.g., *rbkxj*), or auditory forms that begin in a way matching no existing words (e.g., *zlooger*, *eengmov*) allow participants to issue an early response to words based only on partial processing – in the auditory case, for instance, on the existence of some known word beginning with a given initial string. In auditory lexical decision, effects of lexical properties that are observed with a materials set in which nonwords are plausible disappear when all nonwords can be easily rejected (McClennan & Luce, 2005).

The present report describes an auditory lexical decision study called BALDEY (Biggest Auditory Lexical Decision Experiment Yet) which, as its name indicates, is very large by comparison with the average 100- to 200-item protocol. Large data sets from visual lexical decision have been available for some years (e.g., Balota, et al., 2007) and have proven extremely valuable in increasing psycholinguistic knowledge of the parameters that

affect respondents' performance in that task. The auditory task requires considerably greater investment in materials construction, and perhaps for this reason, no such large database for the auditory version of the task has hitherto been compiled; in the study by Luce and Pisoni (1998), for example, which has been considered a large dataset for this task, listeners heard around 300 real and 300 pseudo-words, all of which were monosyllabic. It is our hope that the present data set will, like the extensive data now available from visual lexical decision, lead to a substantial increase in understanding of this useful task.

For all tested items we present reaction times (RTs) and accuracy rates, plus frequency data, as well as overall item durations and a duration measure for each component phoneme in each item. Besides being unusually large and descriptively comprehensive in this way, the data set is also unusually representative in comparison to data from individual experiments. In natural speech, listeners hear a substantial proportion of morphologically complex words. Nevertheless, many auditory lexical decision studies, including most of those cited above, have confined their real-word stimuli to uninflected forms, and even in some cases to uniform structures such as the monosyllables of Luce and Pisoni (1998). Where morphological complexity has been addressed in the experiment (e.g., Baayen, McQueen, Dijkstra & Schreuder, 2003), this has been by means of a direct comparison between forms of the same stem. In our data set, the words vary naturally in morphological structure, and the item-specific information presented includes this structural information too.

In the present report, we describe the construction and collection of the data, and how it can be accessed. To illustrate some of the possibilities opened up by this new dataset, we also present two summary analyses, concerning the sensitivity of the data to the estimation of the point in the word at which recognition may occur, and to different measures of frequency. As can be seen, these analyses concern general properties of the dataset and are not designed to test particular predictions from spoken-word recognition models. Especially given the

phonological and morphological richness of the data, many theoretically driven analyses of the role of different lexical attributes are conceivable.

The data set is publicly available as an ASCII file in which every row represents one trial in the experiment, listing morphological, phonological and acoustic properties of the word presented, the participant's characteristics and the participant's response and response latency. The appendix shows the full list of information currently included in the data file for every trial. This information can be extended, by any users of the database, and we hope that this will happen. In addition, the package includes the audio files of the stimuli with Praat textgrids (Boersma, 2001) providing the HTK phonemic transcriptions (aligned with the signal, see below). The files can be downloaded from http://www.mirjamernestus.nl/Ernestus/Baldey.

## Method

### Participants

Ten male and ten female undergraduate university students took part in the experiment. All were native speakers of Dutch, were aged between 18 and 23 years, and had lived most of their lives in the Dutch province of Noord-Brabant. Four male and two female participants were left-handed. After completion of the full ten sessions of the experiment, participants received 75 euros for their participation.

### Materials

The experiment contained 5541 stimuli: 2780 real words and 2761 pseudo-words.

*Pseudo-words.* To ensure that morphological and phonological structure were balanced across the word and pseudo-word sets, the items were paired such that each pseudo-word was created by changing one or two segments of a real word, leaving affixes intact. For the

resulting pseudo-words, the existing words in the experiment were not always the nearest neighbors (for instance, the pseudo-word *meding* is based on the existing word *lading* 'load', but the existing word *mening* 'opinion', which does not occur in the experiment, is phonologically closer). Since for some words it proved difficult to derive a pseudo-word that had not already been included in the experiment, the number of pseudo-words is slightly lower than the number of real words.

The pseudo-words were constructed to be plausible as lexical candidates (i.e., begin with a phoneme sequence represented in the lexicon) while varying in structure in the same way as the real words; over 60% of them became a pseudo-word only on the final, penultimate or antepenultimate phoneme (respectively 539, 591 and 535 of the 2761 pseudo-words; note that the mean item length in phonemes was 6.8). Examples from the 6-phoneme pseudo-word set are *bewark* (nearest real-word neighbour *bewaren*), *zepels* (*zepen*), and *proemer* (*proef*). Thus no general strategy of superficial word-nonword decision would have been supported.

*Properties of the stimulus set*. Tables 1 to 4 describe the stimulus set. The words represent different categories (differing in word class, morphological structure, the specific affixes, position of stress and number of syllables). Nearly every category contains approximately 40 or 50 words. A category was only incorporated if 40 good representative words could be selected.

As can be seen from the overview in Table 1, relatively few (just over 18%) items are morphologically simple, i.e. have no affixes or have semantically opaque internal structure. Most real stems (1553) occur only once in the stimulus set; however, because Dutch does not contain enough high-frequency stems for stem recurrence to have been avoidable in a set of this size, 698 stems occur between two (458 stems) and seven (2 stems) times (note that the

total number of stem occurrences is greater than the total number of real-word stimuli because compounds have two stems). For instance, the stem *vraag* "ask" occurs as a bare stem (uninflected noun or verb), with the prefixes *over-* (*overvragen* 'to over-demand') and *be-* (*bevraagt* 'questions someone/something') and in the compound *vraagcurve* 'demand curve'). Non-existing stems never recur.

The range of word length was one to five syllables. Of the 2448 polysyllabic real words (with which pseudo-words were paired), 1602 have primary stress on the initial syllable, reflecting (with the monosyllabic pairs) the strong tendency towards initial stress in the Dutch vocabulary (Schreuder & Baayen, 1994). Most syllables are complex. For instance, of the 332 monosyllabic real words, 2.1% consist of a single consonant followed by a vowel and 29.6% consist of a consonant-vowel-consonant string, so that over 68% have at least one consonant cluster, with no fewer than nine words containing five consonants (e.g. *herfst* /hɛrfst/ ’autumn’, trends /trɛnts/ 'trends'). Of the 1219 real bisyllabic words, likewise, 10.8 % consist of only simple (consonant-vowel) syllables (most ending in schwa since the speaker did not realize word-final /n/ after schwa, as is common in Standard Dutch) and 29.0% contain seven consonants or more. The high number of complex syllables is characteristic of the Dutch lexicon and partially results from suffixes such as /s/ (plural) and /t/ (third person singular present tense or past participle marker).

(Tables 1 to 4 about here)

Table 2 shows the number of adjectives, nouns, and regular and irregular verbs without derivational affixes among the real words (and their corresponding pseudo-words), with number of syllables in the words’ stems, the primary stress position and whether the words occur in inflectional forms. Note that derivational affixes may be prefixes or suffixes;

the past participle may be a prefix plus a suffix, and all other inflections are suffixes. In BALDEY (a) inflected adjectives consist of the stem plus /ə/, and are the most frequently used forms of adjectives; (b) the inflectional form used for nouns is the plural; it consists of the stem plus either /ə/ or /s/; (c) four inflectional suffixes for verbs are used, respectively marking (i) third person singular present tense (stem + /t/); (ii) third person plural present tense (stem + /ə/), which is homophonous with the infinitive; (iii) third person singular past tense (stem + /tə/ or stem + /də/ if the verb is regular), which is homophonous with the plural past (see below); and (iv) the past participle (for regular verbs: /xə/ + stem + /t/, unless the stem starts with an unstressed prefix, in which case there is no /xə/).

Table 3 shows the derivational affixes used in the experiment and the numbers of words with each affix. The four prefixes and 12 suffixes are productive and semantically transparent. They are the complete set of affixes for which we could find minimally 40 words that most participants are likely to know. Table 4 shows the number of compounds in the experiment and how often they occurred in inflected form. Most real compounds and pseudo-word compounds are combinations of two existing nouns, with the exception of 142 noun-adjective or adjective-noun combinations. The words with derivational prefixes and the compounds have lemma frequencies of at least 1 in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995).

*Subjective frequency ratings*. To assess participants' likely familiarity with the 1120 single-stem words without derivational affixes, we conducted a rating experiment via the internet. Seventy-five participants (mostly undergraduates) indicated for each word how often they thought an average speaker of Dutch uses the word, on a scale from one (very rarely) to ;seven (very frequently). If they chose "very rarely", they were also asked whether they knew the word and its precise meaning. Most words (821) were known to all participants, while 261 words (e.g. *deun* and *pij*) were unknown to a few (maximally 10 participants). One word

(*ramsj*) was unknown to 46 participants, while the next least well known word was unknown to 36 participants. For 169 words (all unknown to at least some participants), some participants indicated that they knew the word, but did not know its exact meaning.

*Stimulus Recording.* The words were recorded in a soundproof booth, with a sampling rate of 44.1 kHz, by a native female speaker of Standard Dutch, raised in the province of Noord Brabant. She articulated the words carefully. Words ending in *–en* were pronounced as ending in /ə/, as is standard for most Dutch speakers. (Note that in consequence, past tense forms in our stimuli are ambiguous as to number; *wenste* [sing.] and *wensten* [pl.] sound the same. Our stimuli contained no pairs of such homophones.) Mean item duration was 682.8 ms (range 220 - 1347 ms) for words, 698.3 ms (range: 234 - 1352 ms) for pseudo-words.

We aligned the acoustic signal with the phonemic transcription of the word by means of an automatic speech recognizer, HTK (Hidden Markov Toolkit; Young et al., 2006), which received for each item as its input the acoustic signal and the phonemic transcription of that item's citation form. In addition, this recognizer made use of 37 monophone models (32-Gaussian tri-state models), which had been trained on the read-speech component of the Spoken Dutch Corpus (Oostdijk, 2002). Its output has been tested in previous studies. Thus, Pluymaekers, Ernestus, and Baayen ( 2006) showed that, for words produced at slow, medium and fast speech rates, it positions 76 % of the phoneme boundaries less than 20 ms from where a phonetically trained human positioned them. Since the items in our experiment were carefully articulated at a slow rate, the differences with human transcribers are likely to be even smaller. Note, moreover, that differences of this size can also be observed between phonetically trained human transcribers (for an overview see e.g. Ernestus & Baayen, 2011). In the resulting transcriptions, the words and pseudo-words did not differ significantly in mean phoneme duration (words: 105.5 ms; pseudo-words: 105.7 ms; $t(5536.045) = -0.36$, $p >$

0.1). The phonemic transcriptions can be used, among other things, for determining the positions of different types of uniqueness points.

*Stimulus lists.* The 5541 items were pseudo-randomized 20 times, once for each participant, and each randomization was divided into 10 parts, one per session. Each such part contained the same number of words and pseudo-words with a single stem and no derivational affix, the same number of words and pseudo-words with derivational affixes, and the same number of real and pseudo-compounds. Consecutive stimuli in a list did not share either stems or affixes.

**Procedure**

Participants were tested individually in sound-attenuated booths. For every participant, the experiment was divided over ten sessions, which were always one week apart. Each session lasted maximally an hour, and contained four breaks of minimally three minutes.

Participants were instructed to decide as quickly and as accurately as possible for each stimulus whether it was a real Dutch word. The stimuli were presented on the screen of a computer running E-prime (Schneider, Eschman, & Zuccolotto, 2002). The course of a trial was as follows: A star appeared for 300 ms in the centre of the screen, announcing the auditory stimulus. The stimulus was then played over headphones. Participants had four seconds, from stimulus onset, to make their decision. If they did not, a reaction time of 0 was registered. Participants pressed the "yes" button on a button box with their dominant hand, or the "no" button with their non-dominant hand.

**Results**

We collected in total 110,820 responses. The participants chose the wrong answer on 9,852 trials (8.9%); the number of incorrect answers per participant ranged from 184 to 943. The

number of incorrect answers was more than twice as high for the words (7030, 12.6%) as for the pseudo-words (2823, 5.1%), suggesting that participants did not know all the words in the experime nt.

Participants' RTs ranged from 0 ms to 3933 ms, measured from word onset, and from -1279 ms to 3544 ms, measured from word offset. The average RT (measured from word onset) was 1371 ms, with a standard deviation of 603 ms. Only 872 RTs (0.8%) were very short (shorter than 500 ms, measured from word onset, hence likely to result from errors). Analyses of the RT patterns across the ten experimental sessions (linear mixed effect models with the log of the RT as dependent variable, with participant and word as crossed random effects, and with session number and trial number within a session as fixed predictors, see below) showed that participants responded more rapidly the more sessions they had already completed ($\beta$: -0.013, $t(110{,}817) = -19.70$, $p < 0.0001$) and, within a session, the more trials they had already completed ($\beta$: -0.000067, $t(110{,}817) = -5.63$, $p < 0.0001$). This practice effect across the experiment presumably reflects incremental experience with the task and the speaker. Analysis of the accuracy pattern (logistic linear mixed effects models with the same predictors as for the RT analysis) showed that participants made approximately the same number of errors in each session, but their accuracy was slightly higher at the beginning of each session and decreased with every trial ($\beta$: = 0.0003, $z = 3.93$, $p < 0.0001$).

## Illustrative Analyses of the Database

### Analysis 1. What is the word's identification point?

With our first analysis we tried to shed light on the strategies that participants adopted in this study. This provides important information about how to interpret the data set. For this, we investigated the timing of the word-nonword decisions.

The first question we may ask is whether participants always wait until the end of the stimulus item before responding. In the data set as a whole, participants pressed a button (either "yes" or "no") prior to the end of the item on only 3.0% of trials (similarly for real words and pseudo-words). The point at which a response may be held to reflect availability of the full stimulus depends on the hypothesised time needed to initiate muscle movements in this situation; however, it is of interest that the percentage of responses increases steadily with time: within 50ms after offset : 1.3%; between 50 ms and 100 ms after offset: 1.9%; between 100 ms and 150 ms after offset: 2.6%; between 150 ms and 200 ms after offset: 4.1%; and minimally 200 ms after offset: 87.1%.

We can then ask whether, for real words, participants' RTs may be influenced by the position of the phoneme at which the word starts to deviate from other words in the Dutch lexicon. Determining the position of this identification point is, however, not straightforward, since it depends on our assumptions about how participants treat morphologically complex words. We computed the positions of two different identification points. We refer to the first one as the Lemma Identification point (LIP); it is similar to the Uniqueness Point defined by Marslen-Wilson (1980), being the phoneme after which the only remaining lexical candidates are morphological continuation forms of the (prefix plus) stem (see also Balling & Baayen, 2012). An example in our corpus is *bananen* ‘bananas’, with the LIP at the second [n] at which point either the plural form of the stimulus, or its singular *banaan* ‘banana’ are possible, but the competitor *banaal* ‘banal’ is no longer possible (note that this example also works for English). A high correlation between the position of this Identification Point and participants' responses would indicate that participants made their decisions before knowing exactly which word form was presented. The second identification point that we will consider is the phoneme at which the word form can be uniquely identified (for *bananen*, the vowel [ə] constituting the plural suffix; in the English version, the fricative supplying the same plural

information). We will refer to this point as the Form Identification Point (FIP). Although this point is frequently in practice also the end of the stimulus, note that for participants responding when they are sure that stimulus offset has been reached, only post-completion silence supplies such certainty.

We investigated which of these identification points best predicts participants' accuracy and RTs. We compare their predictive powers with that of an identification point always located at word offset. That is, we compared statistical models with as predictor either the duration of the interval between word onset and the end of the phoneme forming the LIP (henceforth LIP interval), the duration of the interval between word onset and the end of the phoneme representing the FIP (henceforth FIP interval), or whole word duration.

**Materials**

*General data set.* We based our analyses on all words bar the four words for which one third of the participants in the rating experiment indicated that they did not know them. For the RT analyses we excluded all incorrect responses (11.4%) and all RTs that were smaller or larger than two standard deviations from the log of the grand mean (this excluded 8678 trials, 1.5%).

*Correlations between the three intervals*. For the words in this data set, word duration correlates better with the word's FIP interval ($r = 0.85$, $t(2774) = 84.0$, $p < 0.001$) than with its LIP interval ( $r = 0.69$ ($t(2774) = 50.4$, $p < 0.001$). This is as expected, since the word's FIP is often located at its final phoneme. The LIP and FIP interval show a correlation of 0.78 ($t(2774) = 66.7$ $p< 0.001$).

Principal component analysis suggests that above all the LIP and FIP intervals differ from each other. The first PCA component explains 85.0% of the variance and represents all

three intervals (correlations between 0.56 and 0.60). The second component, explaining 10.6% of the variance, represents mostly the LIP interval (r = -0.78) and to a much lesser extent the FIP interval (r = 0.15; word duration: r = 0.60). The third component chiefly represents the FIP interval (r = 0.79) and the LIP interval the least (r = -0.27; word duration: r = -0.55).

**Procedure**

*Akaike Information Criteria*. Since the intervals are highly correlated (pair-wise comparisons show correlations between r = 0.69 and r = 0.85; see above), a regression model containing all these intervals as predictors may not reliably show their order of importance (see e.g. Farrar & Glauber, 1967; Montgomery, Peck, & Vining, 2012; Wurm & Fisicaro, 2014), and the order shown for this data set could not reliably serve to predict order in another data set. We therefore constructed one linear mixed effects model (e.g. Baayen, Davidson & Bates, 2008) for each interval separately and compared the models' Akaike Information Criteria (Akaike, 1973).

The Akaike information criterion (AIC) is a measure of the relative quality of a statistical model, for a given set of data, taking into account both the model's goodness of fit of the data and its complexity. The lower the criterion, the better the model. Comparison of the Akaike Information Criteria (AICs) of different models only differing in one predictor will therefore reliably show the order of importance of these predictors. Note that the AICs do not provide information about the relative goodness of models predicting different data sets (e.g., the AICs provide no information about whether a model predicting accuracy shows a better fit with the data than a model predicting RTs).

Not every difference between AICs is meaningful. The formula $\exp((AIC_{min} - AIC_i)/2)$ indicates the probability that the model with $AIC_i$ minimizes the information loss

with respect to the model with $AIC_{min}$. In other words, this formula indicates the likelihood of the model with $AIC_i$ relative to the model with $AIC_{min}$. Since we will be analyzing a relatively large data set, inter-model AIC differences are expected to arise even if the predictors of interest do not substantially differ in their performance. We therefore only consider significant an AIC difference of minimally 14, which implies that the likelihood of the model with $AIC_i$ is maximally 0.001 relative to the model with $AIC_{min}$.

In all linear mixed effects models reported in this article, the predictor of interest (the interval in this analysis and the frequency measure in the second analysis) has a statistically significant effect ($p < 0.01$) according to its t-value. We do not report their coefficients since these are less informative about the models' goodness of fit (and therefore the performance of the predictor of interest) than the AIC itself.

As stated above, this analysis method was chosen because the variables to be compared are highly correlated. However, this method has a further advantage: the inter-AIC differences not only reveal which variable is the best predictor, but also indicate the exact difference in predictivity between the variables. A regression model incorporating all variables simultaneously would not provide this information.

*Combining the intervals.* In addition to comparing the three intervals to each other, we also compared them to predictors representing combinations of them. If a combined predictor outperforms the simple predictors, this may suggest more complex processing than is suggested by the models with simple predictors.

The components of the predictors' PCA discussed above all represent combinations of the predictors (although some components are clearly based more on one interval than another, e.g. PC2 and PC3). The components are by definition not highly correlated and we entered them in the same regression model (that is, we conducted principal component

regression analysis, see, e.g., Jolliffe, 1982; Merz & Pazzani, 1999). If models with one or more of these principal components outperform the models with the single intervals, this would therefore indicate that a combination of the intervals explains the data better than any of the single intervals. We entered all principal components in the first models and removed those that were not significant. We report the analysis with the highest number of principal components with statistically significant effects.

*Statistical modeling.* Two variables known to explain part of the variance in auditory lexical decision data (and also explaining part of the variance in our data) were used as control predictors (i.e. as covariates) in these statistical models: the RT to the previous trial (indicating the participant's local speed) and the number of the trial in the experimental session (controlling for fatigue or learning effects). Word duration and the LIP and FIP intervals were then added to the control models (one for accuracy and one for RT) containing these two control predictors. The control models also contain different intercepts for every word and participant (which means that words as well as participants vary on the two measures). In addition, they contained random slopes for RT to the previous trial by word and participant (which allows the effect of the RT to the previous trial to be different for every word and for every participant). Finally, the RT control model also contained a random slope of trial number by participant (which allowed the effect of trial number to differ per participant). These random effects and random slopes proved to be statistically significant ($p < 0.001$) in ANOVAs comparing models with and without these random effects and slopes.

RTs were measured from word onset. In order to obtain an approximately normal distribution for the RTs, we applied a logarithmic transformation to these RTs, which we then also applied to the RTs to the previous trials. For accuracy, we used logistic regression models with the binomial link function (Jaeger, 2008).

**Results and discussion**

Table 5 shows the AICs for the control models and for the three models incorporating as predictors either word duration, the LIP interval or the FIP interval.

(Table 5 about here)

*Accuracy.* The control model predicting accuracy clearly has a higher AIC (i.e., performs worse) than the other three accuracy models. Thus, addition of a distance measure to some identification point in the word always improves the model. The three accuracy models with such a single distance measure hardly differ from one another in AIC (range: 30283 to 30294). Further, models with measures representing combinations of these intervals (that is, with components of the PCA of the three intervals described above) have AICs in the same range (30283 to 30286). Our data therefore do not provide a decisive answer as to which of these distance measures best predicts accuracy in experiments such as this. This is not unexpected since the participants made few errors, of diverse nature, and the three interval measures are highly related.

*RTs.* The models predicting RTs show a much larger range in AIC (from -4806 to -6931). By far the best model here is the one containing word duration as a predictor. This model as well as the second best model, which contains LIP, outperform the model containing no interval as predictor. This is not the case for the FIP model, which has a higher AIC than the control model.

All single interval models are outperformed when we consider additionally a model containing the first two components of the PCA of the three intervals (as described above): that model obtained an AIC of -7049. As mentioned above, PC1 correlates equally with all three intervals, but PC2 shows the highest correlation with the LIP interval (and word

duration). We conclude that word duration is the best predictor (as shown by the comparison of the models containing single interval predictors), but especially also the LIP interval explains some of the variance (as shown by the model containing PC1 and PC2). Note that this conclusion is also supported by our finding that the LIP model has a substantially lower AIC than the FIP model.

In conclusion, the RT analysis suggests that our participants tended to wait until they had heard the last phoneme in the word before making their decision. However, at least for some words, they started making the decision as soon as they could identify the word's stem.

**Analysis 2. Which word frequency measure captures participants' experience?**

With our second analysis we address a question that is relevant for all future analyses of the dataset (or parts of it). It has long been known that participants' responses in auditory lexical decision experiments are affected by the words' frequencies of occurrence (e.g., Connine, et al., 1990; Dupoux & Mehler, 1990; Luce & Pisoni, 1998; Taft & Hambly, 1986). All analysis of the accuracy and RTs should therefore incorporate lexical frequency as a co-variate, to reduce the variance in the data. The question then is, of course, from which data source these frequencies should be taken.

The design and analyses of many psycholinguistic experiments have been based on the CELEX lexical database, which provides form frequencies from written corpora. Recently, however, Keuleers, Brysbaert & New (2010) showed that participants' responses in a large visual lexical decision experiment correlate more significantly with frequencies calculated from SUBTLEX-NL, an extensive database based on film subtitles, which may be considered a written representation of spoken frequencies. We investigated whether this is also true for auditory lexical decision. Note also that in comparison with the visual lexical decisions compared by Keuleers et al., the items in our experiment show more phonological

variation (the number of syllables ranges from one to five, instead of two) as well as more morphological variation (our experiment also contains, for instance, compounds).

We compared SUBTLEX not only with CELEX but also with two further frequency measures. First, we used frequencies from the Corpus Gesproken Nederlands (CGN; Spoken Dutch Corpus, Oostdijk, 2002). CGN might be expected to outperform SUBTLEX, given that CGN contains many hours of unscripted speech, and may therefore better approach word frequencies as they occur in natural speech. We also examined whether a combination of the CELEX, SUBTLEX and CGN frequency measures may explain the variance in the data better than a single frequency measure. Finally, we investigated the predictive power of the ratings that we obtained in our own internet-based rating experiment.

For CELEX, SUBTLEX and CGN, we tested the predictive power of both word form frequencies and lemma frequencies. Since many of our stimuli were morphologically complex, the relative predictive power of lemma versus word form frequencies may differ (see, e.g., Pinker, 1991).

## Method

### Materials

*General data set.* We compared the predictive power of the frequency measures from CELEX, SUBTLEX and CGN in our lexical decision data across the 1652 real words (with and without derivational affixes) in the experimental data set with form frequencies greater than zero in each of these data bases. This number of words is smaller than the total number of real words in the experiment mainly because CGN only contains 1799 of the words. One of the 1652 words was unknown to 24 participants in the rating experiment. The other words were unknown to maximally 15 participants. Again the analyses of RTs were only based on

correct answers and on RTs that were within two standard deviations of the grand mean. Of the total of 33040 answers, just 8.4% were incorrect.

*Correlations between the word form frequencies*. The three word form frequencies (all log transformed) are highly correlated with one another for this word sample (rs > 0.8, ps < 0.001). PCA revealed that the SUBTLEX form frequency pattern in this sample differs slightly from those of the CELEX and CGN form frequencies. All three frequencies load on PC1 (correlations range between 0.57 and 0.58), which explains 88.5% of the variance. In contrast, PC2, which explains almost 7% of the variance, represents SUBTLEX more (r = 0.81) than CELEX (r = -0.51) or CGN (r = -0.28). PC3 represents CELEX (r = 0.63) and CGN (r = -0.73) more than SUBTLEX (0.13). These PCA components were entered in the regression models presented below to represent combinations of the three form frequency measures.

*Subjective frequency rating*. We investigated the predictive power of the subjective frequency ratings for the subset of 922 real words that occurred in that rating experiment and have form frequencies greater than zero in CELEX, SUBTLEX and CGN (198 words tested in the rating study had no positive frequencies in CELEX, SUBTLEX and CGN). The average per-word rating correlated well with all form frequency measures (between r = 0.73 and r = 0.81, ps < 0.001). A principal component analysis with as input the three form frequencies as well as the subjective frequency rating revealed that rating patterns most closely with CELEX form frequency and least well with SUBTLEX form frequency. For instance, PC2, which explains 7% of the variance, represents rating the best (r = -0.82) and SUBTLEX the least (r = 0.07), while CELEX and CGN show correlations in between (0.54 and 0.17, respectively).

*Correlations between the lemma frequencies.* Lemma frequencies from CELEX, SUBTLEX and CGN were analyzed for the same 1652 real words as in the form frequency comparison. Interestingly, in contrast to the form frequencies, the lemma frequencies (also all log transformed) do not correlate very well. CELEX lemma frequency shows a correlation of 0.09 ($t(1650) = 3.69$, $p < 0.001$) with SUBTLEX lemma frequency and of 0.11 ($t(1650) = 4.60$, $p < 0.001$) with CGN lemma frequency. CGN lemma frequency and SUBTLEX lemma frequency show a higher correlation ($r = 0.42$, $t(1650) = 19.05$, $p < 0.001$).

PCA also shows that CGN lemma frequency and SUBTLEX lemma frequency pattern together. PC1, explaining 49% of the variance, shows correlations of 0.67 and 0.68 with SUBTLEX and CGN, but a correlation of only 0.29 with CELEX. The same holds for PC3, which explains 19% of the variance (SUBTLEX: $r = 0.70$; CGN: $r = 0.71$; CELEX: $r = 0.04$). PC2, in contrast, explaining almost 32% of the variance, represents mostly CELEX ($r = 0.96$, SUBTLEX: $r = -0.24$: CGN: $r = 0.18$). Also these PCA components were entered in the regression models presented below, to represent combinations of the lemma frequency measures.

**Procedure**

*Statistical modeling.* We analyzed how well the different frequency measures correlate with participants' accuracy and RTs, again by comparing the AICs of (logistic) linear mixed effects models. In order to obtain an approximately normal distribution for the RTs, we applied again a logarithmic transformation to these RTs, which we then also applied to the duration of the word and to the RT to the previous trial (which were used as control predictors, see below). RTs were again measured from word onset.

The frequency measures (and their combinations in the form of PCA components) were added to control models containing the same control predictors used in Analysis 1

above. Given the Analysis 1 results, we also added word duration as a control predictor. Note that the control models contain no predictors reflecting morphological properties, since these properties are highly correlated with the frequency measures under investigation (e.g., the log of the SUBTLEX form frequency has a mean of 5.0 for our words with one stem and no derivational affixes, of 3.6 for words with derivational affixes, and of 2.7 for the compounds).The control models also contain statistically significant intercepts for word and participant, random slopes for RT to the previous trial by word and participant, and for word duration by participant. In addition, the RT control model also contains a random slope of trial number by participant.

**Results and discussion**

Table 6 shows the AICs for the control models (without frequency measure), for the models with the three form frequencies and for the models with the three lemma frequencies, for the dataset of 1652 words. We first discuss the results for the objective and subjective word form frequencies, for the accuracy and RT data in parallel; discussion of the lemma frequencies follows.

(Table 6 about here)

*Word form frequencies.* The presence of a predictor reflecting form frequency improves the model more if this predictor is based on SUBTLEX than if it is based on CELEX. Since Keuleers et al. (2010) obtained the same results for visual lexical decision, the explanation cannot be found in how well the modalities represented by the data bases match the modality in our lexical decision experiment (CELEX based on written language versus SUBTLEX based, albeit indirectly, on spoken language). Following Keuleers et al., we propose that SUBTLEX outperforms CELEX because it better represents our participants' experience with their native language. CELEX is based on carefully edited written texts,

while our participants are more familiar with the rather more informal language which constitutes a good proportion of the SUBTLEX corpus.

CGN contains recordings of completely spontaneous casual speech, which of course represent informal language even better than SUBTLEX does. In addition, however, CGN contains formal speeches, lessons and news bulletins, which are more similar to written language. This mixture of speech styles, but of course also the smaller size of CGN (9 million words, compared to SUBTLEX's 44 million words) may explain why the SUBTLEX form frequencies outperform not only CELEX but also the CGN form frequencies.

*Combination of the word form frequencies.* We investigated whether SUBTLEX form frequency also better accounts for the data than a combination of the three form frequency measures. We therefore added to the control models the components of the PCA of the three word form frequencies (described above). We ran models with all PCAs and with just subsets of PCAs (following the procedure described above). The resulting models did not have significantly lower AICs (accuracy: minimally 16324; RT: minimally -6396) than the model with only SUBTLEX form frequency.

*Subjective frequency rating.* For the subset of 922 real words that were also incorporated in the rating study, the models with SUBTLEX form frequencies (AIC accuracy: 16326; AIC RTs: -2924) also outperform the models with subjective frequency rating as predictor (AIC accuracy: 16511; AIC RTs: -2509). These results are in line with Brysbaert & Cortese (2011), who claimed that objective frequency measures outperform subjective frequency measures if these objective measures well reflect the participants' language experience.

We then examined whether the subjective frequency ratings may account for some of the variance that is not accounted for by the objective word form frequencies. For this, we

compared two types of models for this subset of 922 real words: models with components of a PCA of the three objective word form frequencies and models with components of a PCA of the three word form frequencies and rating (see the PCA analyses discussed in the Materials section).On the accuracy measure, the latter type of model (AIC of best accuracy model, with all four components: 10219; ) outperforms the first type of model (AIC of best accuracy model, with PC1 and PC2: 10338). The difference between the two types of models for the RT shows the same pattern (AIC of best RT model, with all four components: -2745 versus AIC of best RT model with PC1 only combining objective frequency measures: -2735) but is not significant according to our criterion formulated above. Together, the analyses suggest that the subjective frequency ratings indeed contain relevant information that is not captured by one of the word form frequencies, not even by SUBTLEX.

*Lemma frequencies*. Our conclusion that of the three objective data bases, SUBTLEX best accounts for our participants' response patterns is supported by the predictive powers of the three lemma frequencies for accuracy (see Table 6): again the frequency measure based on SUBTLEX outperforms the measures derived from CELEX and CGN.

The situation is different for the models predicting RTs: the best performing model incorporates lemma frequency from CELEX rather than from SUBTLEX or CGN. This result strongly suggests that CELEX better represents the variety of words (and their frequencies) that a participant knows with a certain stem than the corpora that well represent informal speech. Informal speech is typically relatively poor in number of word types, and our participants may have learnt the majority of the word types that they know from written texts. CELEX, which is based on written text, may therefore better reflect participants' knowledge of word lemmas.

*Combination of the lemma frequencies.* Next, we investigated whether a combination of the three lemma frequencies has a better predictive power than any of the lemma frequencies separately. We incorporated the components of the PCA of the three lemma frequencies described above to the control models. On the RT measure, the model incorporating just CELEX lemma frequency remains the best model (the lowest AIC, obtained with all three PCA components, is -6396). CELEX lemma frequency by itself is clearly the best predictor.

In contrast, on the accuracy measure, the best model contains components of the PCA (AIC of the best model, with all three components: 16325). These results thus suggest that a combination of lemma frequencies outperform SUBTLEX lemma frequency as predictor for accuracy. This brings the results for accuracy more in line with the results for RT: both accuracy and RT are well predicted by CELEX. Note that given the low error rate (8.4%) and the possibility that errors could be of diverse kinds, the results for accuracy may be less conclusive.

*Word form versus lemma frequencies.* Finally, Table 6 shows that for both the accuracy and RT data, SUBTLEX and CGN word form frequencies are better predictors than the corresponding lemma frequencies. This suggests that participants' recognition of morphologically complex words was based on the word forms themselves rather than their stems. This finding is in line with earlier research showing that Dutch listeners recognize morphologically complex words via their whole forms, rather than their stems (e.g. Baayen et al., 2003).

CELEX lemma frequency and form frequency, in contrast, are equally good predictors of accuracy, while CELEX lemma frequency outperforms CELEX form frequency for RT. This difference between CELEX on the one hand and SUBTLEX and CGN on the

other may be related to our finding that lemma frequency predicts RT most accurately if derived from CELEX. Which frequency measure performs best can therefore depend on the corpus the frequencies are extracted from.

**General Discussion**

BALDEY, the Biggest Auditory Lexical Decision Experiment Yet, provides experimenters in the field of spoken-word recognition with a resource allowing many research questions to be posed. Through the initial analyses reported in this contribution, it has also already supplied some handy guidelines for the design and evaluation of studies with this task.

As described in the introduction, the listener's task in recognising spoken words is to discard potential alternative interpretations of what is being presented and settle, as rapidly as possible, upon the selection of known words actually corresponding to the utterance being heard. The operation of word recognition proceeds in the same way irrespective of the particular language in which the utterance is couched, as it follows necessarily from the structure of vocabularies. Thus the listener's task is the same in a polysynthetic language where highly complex words are composed of elements that never stand alone, in a language such as those of China, where all forms are simple and may stand alone, or in a language that makes words of intermediate complexity combining stand-alone elements with bound morphemes (as in English, or in the language used in BALDEY, Dutch). In all languages, multiple possible lexical interpretations become available, but the listener is able to discard most of them rapidly and achieve recognition efficiently.

The two general and theory-neutral issues which we investigated in the first use of the BALDEY data set, and report in this paper, extend our knowledge of how the auditory lexical decision task is performed by listeners, and thereby motivate some suggestions for its useful deployment in future. The first issue concerned the task-specific yes-no decision required of

listeners, and the timing of this decision. In natural speech situations, listeners may usually safely assume that speech signals consist solely of real words, and they may also draw upon accumulating evidence from interpretation of the utterance so far, and from knowledge of the discourse context, to inform their choice between alternative potential interpretations supported by the acoustic input. Neither of these statements is true of an auditory lexical decision experiment. The task presents input which has a roughly equal likelihood of being a real word or of being nonsense. And in addition, there is no discourse context or other probabilistic evidence to constrain decisions.

Accordingly, researchers using the auditory form of lexical decision have often assumed that listeners will not accept a word as real until they are sure that they have heard the whole word (i.e., that the word is not going to continue with some further phoneme or phonemes that would render it a nonword). The results of our analyses of the BALDEY data suggest that participants did indeed respond in a way suggesting such caution. This is clear from the high number of responses (97%) issued after stimulus offset. Furthermore, we calculated for each real word the point at which it became a unique lemma and could only become itself or a related suffixed form (LIP), as well as the point at which it was a unique word form (FIP). Neither of these measures could as strongly predict our participants' responses as the duration of the whole stimulus item. In other words, in the large majority of cases in this set of stimuli (deliberately chosen to represent the structural complexity of lexical items encountered in natural speech), responses were made only once the complete stimulus had been heard.

Note that we did not analyze the effect on responses to pseudowords of the point at which such a stimulus could be definitively distinguished from real Dutch words (Nonword Identification Point, or NIP, to adapt Marslen-Wilson's terminology). We have no doubt that

NIP would be related to response patterns, since this has been shown in some of the earliest literature with this task (Marslen-Wilson, 1980).

The early spoken-word recognition literature predates the availability of computer-searchable lexical databases. Models proposed in the 1970s (e.g., Cole & Jakimik, 1978; Marslen-Wilson & Welsh, 1978) made much of the temporal nature of speech recognition and of the fact that later parts of spoken words were sometimes redundant. When lexical databases became available in the 1980s, however, calculations (e.g., by Luce, 1986) soon revealed that such redundancy was generally only to be found in longer words, while the greater part of the vocabulary consisted of shorter words (which also tended to occur more frequently). Moreover, longer words very often had shorter words embedded within them, so that the first word activated by the speech signal *catalogue* is the unrelated form *cat* (note that this example also works for Dutch *catalogus*/*kat*). Studies with lexical decision had, as already noted, shown that information later than the uniqueness point of the word affected decisions (Goodman & Huttenlocher, 1988; Taft & Hambly, 1986), and research with incremental presentation of words also showed that recognition prior to the word's end was rare (Bard, Shillcock & Altmann, 1988; Grosjean, 1985). Strictly left-to-right models of spoken-word recognition were thus replaced in the 1980s and 1990s by models involving competition between simultaneously activated word forms (Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986; Norris, 1994). Multiple activation and competition have typically been tested and demonstrated with tasks other than auditory lexical decision.

Using the task with nonwords, however, reveals clear evidence of competition (Marslen-Wilson & Warren, 1994; McQueen, Norris & Cutler, 1999; Taft, 1986). The BALDEY evidence confirms the presence of competition in performance of the task, and further supports the computation of all response times from item offset. This reflects the point at which listeners issue responses, as well as eliminating irrelevant variance caused by simple

effects of item duration. Once the acoustic evidence is in, a decision may be made and a response issued. At that point, the speed and accuracy with which this happens will be a function of many factors, some of which should of course be the factors manipulated in the particular experiment.

One factor that has a strong effect on participants' responses in any word processing task is, of course, their familiarity with the words they are processing. This is typically estimated by consulting counts of word occurrence frequencies. The second issue that we addressed in the BALDEY data was the relative ability of different frequency measures to account for the patterns revealed in the data. We observed that the response patterns (both in accuracy and RTs) were better captured by form frequencies in a very large database compiled from film subtitles (SUBTLEX) than by frequencies of forms in written text (CELEX) or in spoken natural communication, both spontaneous and rehearsed (CGN), or by subjective ratings collected in an online experiment with (a subset of) the BALDEY stimuli. We suggested several reasons why SUBTLEX form frequencies should have provided a better account of our data: First, it is by far the largest corpus from which frequencies have been calculated; second, it contains speech that is (or at least is supposed to be) largely natural conversation. Since natural conversations certainly constitute the primary source of input on which participants' listening experience will have been based, the SUBTLEX corpus is putatively closer to the source of form frequency effects in these listeners' processing.

For lemma frequencies, in contrast, our participants' RTs were better predicted by the frequency values given in CELEX than by those in the other sources. Since CELEX frequency is based on written corpora, it may be that participants' responses not only reflect knowledge about the lemma's existence, but also reflect the fact that a significant part of their experience of these lemmas is from reading.

The results were in general clearer in the RT than in the accuracy analyses. The most likely reason for this is that the latter analyses were highly skewed, comparing 88.6% (correct responses) to only 11.4% (wrong decisions plus potentially anticipatory decisions), which leads to a relatively low statistical power. Moreover, there may be several reasons why participants erroneously classified a real word as a non-word, including of course being unfamiliar with the word.

In accord with prior findings, form frequency had a stronger predictive power than lemma frequency for both the RT and the accuracy measure in our data, where the frequency measures were derived from SUBTLEX or CGN. For CELEX-derived measures, we found either no difference between form and lemma frequency (accuracy) or lemma frequency outperforming form frequency (RT). This shows how sensitive this type of comparison is to the corpus from which frequencies are derived. It is not surprising that among the different lemma frequencies, the one based on CELEX outperforms form frequency because this lemma frequency predicted RT best.

The better performance of form frequency compared to lemma frequency from SUBTLEX and CGN suggests that listeners' decisions reflected access to the word's exact form, rather than just its stem. This is fully consistent with the results of our first analysis, showing that listeners' decisions were overwhelmingly issued after the end of the stimulus word. Note that spoken-word recognition in natural speech contexts requires morphological, syntactic and discourse processing which may render such form-specific judgments necessary.

Nonetheless, it is worth pointing out that all the frequency measures we included in this analysis were significantly predictive of the response patterns we found, and indeed nearly all were highly correlated with one another. Further, several combined analyses

revealed that each of the frequency measures (from the different sources) captured some further variance in comparison with the most strongly predictive measure. These analyses thus suggest that in general experimental practice any measure of frequency based on a reasonably extensive underlying sample will serve to tap into frequency effects in an auditory lexical decision data set. The bigger the corpus in question, of course, the more reliable its predictions of effects will be. For researchers who are chiefly concerned with basic word knowledge, lemma frequencies based on written language will be the count of choice; for researchers particularly interested in speech uptake, frequency counts based on a spoken language source (such as the SUBTLEX corpus) will be preferable.

In conclusion, we greatly look forward to the many further uses to which the BALDEY data set will be put in the future.

**References**

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & B.F. Csaki (Eds.), *Second International Symposium of Information Theory* (pp. 267–281). Budapest: Academiai Kiado.

Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419–439.

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., McQueen, J. M., Dijkstra, T., & Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 355–390). Berlin: Mouton de Gruyter.

Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database* (Release 2) [CD ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Balling, L.W., & Baayen, R.H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition,* 125, 80–106.

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

Bard, E.G., Shillcock, R.C. & Altmann, G. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics, 44,* 395–408.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.

Brysbaert, M., & Cortese, M.J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology*, *64*, 545–559.

Cole, R.A. &, Jakimik, J. (1978). Understanding speech: How words are heard. In G. Underwood (Ed.), *Strategies of Information Processing* (pp. 67–116). London: Academic Press.

Connine, C.M., Mullennix, J, Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 1084–1096.

Dupoux, E., & Mehler, J. (1990). Monitoring the lexicon with normal and compressed speech: Frequency effects and the prelexical code. *Journal of Memory and Language, 29*, 316–335.

Ernestus, M., & Baayen, R.H. (2011). Corpora and exemplars in phonology. In J. Goldsmith, J. Riggle, & A. Yu (Eds.), *The Handbook of Phonological Theory* (2nd ed.), (pp. 374–400). Chichester, West Sussex: Wiley-Blackwell.

Gaskell, M.G., & Marslen-Wilson, W.D. (1997). Integrating form and meaning, A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613–656.

Farrar, D.E., & Glauber, R.G. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics, 49*, 92–107.

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

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28,* 501–518.

Goodman, J. C., & Huttenlocher, J. (1988). Do we know how people identify spoken words? *Journal of Memory and Language, 27,* 684–698.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics, 38,* 299–310.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language,* 59, 434–446.

Jolliffe, I.T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 31,* 300–303.

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

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). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1–36.

Marslen-Wilson, W. D. (1980). Speech understanding as a psychological process. In J. C. Simon (Ed.), *Spoken language generation and understanding* (pp. 39–67). Dordrecht: Reidel.

Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101,* 653–675.

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

McClelland J. L., & Elman J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 306–321.

McQueen, J. M. (1996). Word spotting. *Language and Cognitive Processes, 11*, 695-699.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. In M.G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 37–53). Oxford: Oxford University Press.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20,* 621–638.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1363–1389.

Merz. C.J., & Pazzani, M.J. (1999). A principal components approach to combining regression estimates. *Machine Learning* 36, 9–32.

Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012). *Introduction to linear regression analysis* (Vol. 821). Hoboken, NJ: John Wiley.

Norris, D. G. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21,* 1209–1228.

Oostdijk, N. (2002). The design of the Spoken Dutch Corpus. In P. Peters, P. Collins, & A. Smith (Eds.), *New frontiers of corpus research* (pp. 105–112). Amsterdam: Rodopi.

Pinker, S. (1991). Rules of language. *Science, 253*, 530–535.

Pluymaekers, M., Ernestus, M., & Baayen, R.H. (2006). Effects of word frequency on the acoustic durations of affixes. In *Proceedings of Interspeech 2006 - ICSLP* (pp. 953–956).

Schneider, W., Eschman, A., & Zuccolotta, A. (2002). *E-prime User's Guide*. Pittsburgh: Psychology Software Tools.

Schreuder, R., & Baayen, R. H. (1994). Prefix stripping re-revisited. *Journal of Memory and Language, 33*, 357–375.

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 131–157). Hillsdale, NJ: Erlbaum.

Slowiaczek, L.M., & Pisoni, D.B. (1986). Effects of phonological similarity on priming in auditory lexical decision. *Memory and Cognition*, *14*, 230–237.

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

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

Tanenhaus, M.K., & Spivey-Knowlton, M.J. (1996). Eye-tracking. *Language and Cognitive Processes*, *11*, 583–588.

Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 98–108.

Wurm, L.H., & Fisicaro, S.A. (2014). What residualizing predictors in regression analyses does (and what it does *not* do). *Journal of Memory and Language, 72*, 37–48.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... Woodland, P (2006). *The HTK book* (v3. 4). Cambridge: Cambridge University Press.

Zwitserlood, P. (1996). Form priming. *Language and Cognitive Processes*, *11*, 589–596.

Number of words: 9791 (only main text and references, excluding abstract, acknowledgements, appendix and figures)

Appendix

Information included in the BALDEY datafile.

- The participant identifier (subject)
- The number of the session in the experiment (session)
- The number of the trial in the session (trial)
- The participant's age (age)
- The participant's gender (gender)
- Whether the participant is left or right handed (hand)
- Where the participant was born (always a region in the Netherlands)
- Whether the participant speaks a dialect and if so from which location in the Netherlands (dialect)
- The participant's highest school degree (diploma): Either participant's level of high school (with vwo being higher than havo) or at least first year of university finished (propedeuse)
- The stimulus (word)
- The word status of the stimulus (word_status)
- Whether the participant classified the stimulus correctly (response)
- Reaction time measured in ms from word onset (RT)
- The stem of the stimulus (stem)
- Reaction time in the previous trial measured in ms from word onset (RTprev)
- Phonemic transcription of the stimulus (transcription)
- Number of phonemes in the stimulus (Nphonemes)
- Whether the stimulus is an adjective, a noun or a verb (word_class)
- If the stimulus is a verb, whether its inflection is regular (regularity)
- Whether the stimulus is an inflected (pseudo) word (inflected)
- The phoneme representing the lemma identification point (lip)

- The position of the offset of the phoneme representing the lemma identification point in the stimulus' wave form, measured from word onset (lip.ms)
- The phoneme representing the form identification point (fip)
- The position of the offset of the phoneme representing the form identification point in the stimulus' wave form, measured from word onset (fip.ms)
- Number of letters in the orthographic representation of the stimulus (Nletters)
- Number of syllables in the stem (Nstem_syllables)
- Number of syllables in the word (Nword_syllables)
- Phonemic transcription of the syllable with primary stress (stressed_syll)
- Whether the initial syllable is stressed (initial_stress)
- Whether the final syllable is stressed (final_stress)
- Whether the stimulus is morphologically underived, derived or represents a compound (morph_classification)
- Duration of the stimulus in ms (word_duration)
- For verbs, the tense of the verb form (tense)
- For verbs, the number of the verb form (number)
- For verbs, the person of this verb form (person)
- For real words, the form frequency taken from CELEX (CELEX_form_freq)
- For real words, the lemma frequency taken from CELEX (CELEX_lemma_freq)
- For real words, the form frequency taken from CGN (CGN_form_freq)
- For real words, the lemma frequency taken from CGN (CGN_lemma_freq)
- Rating as obtained in the web experiment (rating)
- Number of participants in the web experiment who classified the word as "unknown" (word_unknown)
- Number of participants in the web experiment who indicated they did not know the meaning of the word (meaning_unknown)

- Whether the word contains a derivational affix and if so which (affix)
- For compounds, the first stem (stem1)
- For compounds, the second stem (stem2)
- For compounds, the word classes (noun versus adjective) of the two stems
- For compounds, the first stem's form frequency taken from CELEX (CELEX_form_freq_stem1)
- For compounds, the first stem's lemma frequency taken from CELEX (CELEX_lemma_freq_stem1)
- For compounds, the second stem's form frequency taken from CELEX (CELEX_form_freq_stem2)
- For compounds, the second stem's lemma frequency taken from CELEX (CELEX_lemma_freq_stem2)

Table 1. Overview of morphological structure and length of items in the stimulus set, with real-word examples of each type between brackets.

| | Real words | Pseudo-words | Number of syllables (mean and range ) |
|---|---|---|---|
| Morphologically simple (no affixes, or opaque structure) (*pover*) | 511 | 500 | 1.7 (1 - 3) |
| One stem, one inflectional suffix (*katten*) | 609 | 613 | 2.0 (1 - 3) |
| One stem, one derivational affix (*kreupelheid*) | 770 | 723 | 2.5 (1 - 4) |
| One stem, two affixes (one derivational, one inflectional) (*beklaagde*) | 370 | 407 | 2.9 (2 - 5) |
| 2-stem compounds (*haarfijn*) | 375 | 414 | 3.1 (2 - 4) |
| 2-stem compounds with inflectional suffix (*dasspelden*) | 145 | 104 | 3.1 (2 - 4) |

Table 2. Real words and pseudo-words without derivational affixes in the experiment, as a function of the number of syllables in the stem, the position of stress, and the presence of an inflectional affix ("3rd ps. sing. present": third person singular present tense).

| Word type | syllables in stem | Position of stress | Inflection | Real word total | Pseudo-word total |
|---|---|---|---|---|---|
| Adjectives | 1 | Initial | No | 40 | 40 |
| | | | Yes | 40 | 40 |
| | 2 | Initial | No | 40 | 40 |
| Nouns | 1 | Initial | No | 126 | 118 |
| | | | Plural | 124 | 131 |
| | 2 | Initial | No | 75 | 75 |
| | | | Plural | 75 | 75 |
| | | Final | No | 50 | 49 |
| | | | Plural | 50 | 51 |
| | 3 | Initial | No | 40 | 40 |
| | | Final | No | 50 | 50 |
| Regular Verbs | 1 | Initial | No | 40 | 38 |
| | | | 3rd ps. sing. present | 40 | 40 |
| | | | Plural present | 80 | 78 |
| | | | Simple past | 40 | 40 |
| | | | Past participle | 40 | 39 |
| | 2 | Initial | No | 50 | 50 |
| Irregular verbs | 1/2 | Initial/Final | Simple Past | 80 | 79 |
| | | | Past participle | 40 | 40 |

Table 3. Real words and pseudo-words with derivational affixes in the experiment, as a function of the number of syllables in the stem, and the presence of an inflectional affix ("3rd ps. sing. present": third person singular present tense; "past part." : past participle; "pl. present": plural present).

| Affix | Syllables in stem | Inflection | Real word total | pseudo-word total |
|---|---|---|---|---|
| +*achtig* [ɑxtəx] | 1 | No | 40 | 39 |
| +*baar* [bar] | 1 | No | 40 | 40 |
| *be*+ [bə] | 1 / 2 | No | 40 | 39 |
| | | 3rd ps. sing. present / simple past / past part. | 40 / 40 /40 | 40 / 40 /40 |
| +*elijk* [ələk] | 1 | No | 40 | 40 |
| +*er* [ər] comparative | 1 / 2 | No | 40 / 40 | 38 / 40 |
| | 1 | Yes | 40 | 41 |
| +*er* [ər] agens | 1 | No | 50 | 49 |
| | | Plural | 50 | 48 |
| +*erig* [ərəx] | 1 | No | 40 | 40 |
| +*erij* [ərɛi] | 1 | No | 40 | 38 |
| +*heid* [hɛit] | 1 / 2 | No | 40 / 40 | 40 / 40 |
| +*ig* [əx] | 1 /2 | No | 40 / 40 | 41 / 39 |
| | 1 | Yes | 40 | 39 |
| +*loos* [los] | 1 | No | 40 | 39 |
| *ont*+ [ɔnt] | 1 | No | 40 | 40 |
| | | Plural | 40 | 40 |
| *over*+ [ovər] | 1/2 | No | 40 | 40 |
| +*schap* [sxɑp] | 1/2 | No | 40 | 40 |
| *ver*+ [vər] | 1/2 | No | 40 | 40 |
| | | 3rd ps. sing. present / pl. present / simple past | 40 / 40 /40 | 40 / 40 / 40 |

Table 4. Number of existing compounds and pseudo compounds in the experiment, broken by the word type of the first part, the word type of the second part, their numbers of syllables, and the presence of inflection.

| Initial part: word class/ syllable count | Second part: word class/ syllable count | Inflection | Real word total | Pseudo-word total |
|---|---|---|---|---|
| Adjective 1 | Noun 1 | No | 40 | 40 |
| Noun 1 | Adjective 1 | No | 31 | 31 |
| Noun 1 | Noun 1 | No | 50 | 48 |
| | | Plural | 50 | 50 |
| Noun 1 | Noun 2 | No | 55 | 55 |
| | | Plural | 55 | 54 |
| Noun 2 | Noun 1 | No | 79 | 80 |
| | | Plural | 40 | 40 |
| Noun 2 | Noun 2 | No | 120 | 120 |

Table 5. AICs of the statistical models for accuracy and reaction times that incorporate the duration of the interval to an identification point in the word, and of their control models (without any interval).

| Duration measure | AIC Accuracy | AIC Reaction times |
|---|---|---|
| None (control model) | 30325 | -5430 |
| Word duration | 30283 | -6931 |
| LIP interval | 30294 | -6075 |
| FIP interval | 30289 | -4806 |

Table 6. AICs of the statistical models for accuracy (a) and reaction times (b) that incorporate frequency measures and of their control models without any frequency measure

| (a) Accuracy | | |
|---|---|---|
| No frequency measure | 16513 | |
| | form | lemma |
| CELEX frequency | 16456 | 16463 |
| SUBTLEX frequency | 16326 | 16379 |
| CGN frequency | 16384 | 16463 |
| | | |
| (b) Reaction times | | |
| No frequency measure | -6612 | |
| | Form | Lemma |
| CELEX frequency | -6674 | -6723 |
| SUBTLEX frequency | -6712 | -6660 |
| CGN frequency | -6686 | -6311 |