What should be the Place of the Normative Database in Speech Perception Research?

Bernadine Cocks, Jana Smith, and Graham A. Jamieson

University of New England
bcocks2@myune.edu.au

To assess the practice of using normative database values to control for confounds in English speech perception research, six separate surveys gathered normative data from a contemporary Australian population on a corpus of 140 English language words. After averaging across participants, these individual word values were then compared with values retrieved from CELEX, the BNC and the MRC. Contrary to predictions, there were significant correlational differences between all four sources for familiarity/frequency, as well as significant mean corpus differences between the current study and the MRC for concreteness and imageability. Furthermore, significant differences were found between written and spoken presentations in both the current study’s surveys and the BNC. This suggests that although the use of normative database values may be common, the practice should be approached with a degree of caution.

Key words: CELEX, MRC, BNC, SUBTLEX, speech

Introduction

Controlling for the effects of potentially confounding variables is standard practice in scientific investigation. From a language research perspective, such variables can include word length, phonotactic probability, neighbourhood frequency, familiarity and concreteness (e.g. Arciuli & Slowiaczek, 2007; Pulvermüller, Lutzenberger & Preissl, 1999; Cortese & Fugett, 2004; D’Angiulli, 2004). While physical attributes such as word length are easily measured and accounted for, more cognitive features such as familiarity and concreteness are not; instead, degrees of difference
are controlled for through the use of normative data values (e.g., Ogawa & Inamura, 1974; Liu, Shu & Li, 2007). As many language researchers have neither the time nor the finances to gather such normative data from a prospective population, there appears a tendency to rely on published databases to obtain normative values. For English language researchers, such databases include CELEX (Baayen, Piepenbrock, & Gulikers, 1995), the British National Corpus (BNC; Aston & Burnard, 1998), Subtlex (Brysbaert & New, 2009), the Corpus of Contemporary American English (COCA; Davies, 2009), and the Medical Research Council Corpus (MRC; Coltheart, 1981), to name but a few.

This reliance on normative database values appears especially commonplace (e.g., Gonsalves & Paller, 2000; Mattys & Samuel, 2000; Helenius, Parviainen, Paetau & Salmelin, 2009); however, there remain small but nagging doubts over this practice. There are a variety of reasons for this. Firstly, a number of these databases derive normative values that reflect frequency of word appearance in printed media such as newspapers and magazines and not spoken language usage. Although the exact nature of the relationship between speech and literacy remains at the heart of an intense debate (for example, see Hauser, Chomsky & Fitch, 2002), there is a growing body of empirical evidence (for example, see Schlagger & McCandliss, 2007; Preston et al., 2010) along with reasonable theoretical argument (for example, see Olsen, 1998) to suggest that despite speech and literacy being closely related, there are fundamental differences between these processes. Until both the differences and similarities between speech and literacy can be more clearly elucidated, using normative values derived from printed media may be appropriate for literacy research paradigms; their use, however, in spoken language research paradigms, especially speech perception studies (e.g. Pulvermüller et al., 1999; Longe, Randall, Stamatakis & Tyler, 2007), leaves both results and interpretations open to question.

Although there are a number of on-going projects attempting to provide spoken language normative values, these too are problematic. For example, the BNC offers spoken word frequency values for 10 million English language words. To obtain these values, the casual speech of 127 participants was recorded over a three day period and frequency values
What should be the Place of the Database? assigned according to the production of individual words. While these normative values may thus accurately capture spoken word production frequencies, humans are able to comprehend substantially greater numbers of words than they produce in casual conversation (Bishop & Snowling, 2004). Furthermore, as demonstrated by the obvious differences in the physical manifestations of Wernicke’s aphasia and Broca’s aphasia, there is direct physical evidence that comprehension and production, like speech and literacy, are related although different processes (Zurif, Swinney, Prather, Solomon & Bushell, 1993). Allowing for this established difference between production and comprehension, the use of speech production frequency values sourced from the BNC for speech comprehension paradigms may not be appropriate.

Similarly, the Subtlex database sources its values from spoken language production, specifically subtitles associated with electronic media (e.g. film and television). Again, however, the values reflect production (frequency) rather than comprehension (familiarity) making such values a questionable choice for speech perception studies. There is also the issue of language style with this database. While screen and script writers attempt to mimic natural speech, any scripted dialogue conforms to the style requirements of the individual production or medium thus is formally constrained by that style; that is, it is not natural (Field, 1998). Given also that electronic media writers are often taught to make their dialogue “realer than real” (personal communication), the use of such electronic sources to gain normative values again appears questionable. Even less formal contexts, such as ad lib interviews or debates, remain constrained by the artificiality of the context and style. As a result, Subtlex is a major step forward for language research in general, but remains problematic for speech perception paradigms.

From a different perspective, spoken language use is known to evolve and change over time (Senghas & Coppolla, 2001). The word gay, for example, was defined in 1951’s Highroads Dictionary (p. 186) as “lively, full of fun, dressed in bright colours”. By 2004 the Australian Concise Oxford Dictionary 4th Edition (Moore, p 577) defined gay as “homosexual…intended or used by homosexuals…of or relating to the homosexual community”. In 2013, however, gay is quickly losing its sexual connotations and amongst Australian adolescents is often used to indicate that someone
or something is not desirable (for example, “That is just so gay” means “That is uncool/nerd-like/geeky”). Such changes can also go beyond relatively simple lexical variations like this, with entirely new languages able to develop in as little as two decades (Senghas & Coppolla, 2001). With regards to the various normative databases available, several appear close to if not past their use-by date. For example, CELEX was last revised in 1995, the BNC in 1993 and the MRC in 1981. As a result, the values published by these databases may well have reflected language use at the time they were compiled, but may not be accurate reflections of language use today.

Another possible criticism of normative database value use is that such norms were, and remain, primarily calculated using Northern hemisphere data, especially American and British English. Although Australia, New Zealand, the United States of America, Canada and Great Britain may all speak English, regional dialectal differences are obvious to even the casual observer. This is not simply confined to differences in pronunciation (e.g., /glæs/ versus /glæs/), but also to stress patterning (“oreGANno” versus “oREGano”) and lexical differences (e.g., the verb “to root” brings somewhat different mental imagery to mind depending on whether you are Australian or American). As a result, the use of normative database values derived from Northern hemisphere English language use may not be an accurate reflection of Australasian English language use, again making the use of such databases in Australasian language research potentially questionable.

Ease of access and use can also restrict a normative database’s practical applicability to language research. For example, the COCA provides free access to the top 5 000 most frequent words. Unfortunately, this includes function words (e.g. “the” and “a”), conjunctions (e.g. “and”), pronouns (e.g. “I”), numerous adjectives and adverbs, as well as varying syllable lengths (from one to three). If (as was the case with two of the current authors), stimulus selection needed to be determined by syllable length and grammatical class type, the amount of free access data available through the COCA would be severely restricted. Given that there are no-charge alternatives, it is understandable that financially-challenged researchers, irrespective of their geographic location or English dialect, would seek normative values from free-to-use sources.
Finally, none of the databases currently available provide a complete range of normative data values for any linguistic feature other than frequency. Although the MRC does provide some normative data values for concreteness and imageability, the list is far from complete. Thus, like the COCA, this severely restricts the number of individual words which can be chosen for use as experimental stimuli which, in turn, impacts on the ability to randomly select stimulus words. As random selection lies at the heart of scientific experimentation, and is a basic assumption of most statistical analyses (e.g., t-tests, ANOVA) this randomness restriction may be more than a mere annoyance in paradigm planning. It also means that researchers are unable to plan their research from certain viewpoints (e.g., grammatical class); instead, they must see what words have values available, then reverse-engineer their designs.

The degree to which any paradigm is potentially confounded by such issues is very much paradigm-dependent; however, designs involving speech perception appear at greatest risk due to normative databases reflecting production rather than comprehension values. This is of particular concern when the experiment involves electroencephalography (EEG) or magnetoencephalography (MEG) where the relatively precise millisecond (ms) timing or locations of speech comprehension/perception neural networks are the subject of investigation. Take, for example, Pulvermüller et al. (1999) who used an event-related potential (ERP) EEG paradigm, to evidence a neural noun/verb double dissociation in less than 300 ms of word stimulus presentation. This dissociation, the authors suggested, involved processing of the two grammatical class types by two different functional neural networks; nouns, it was proposed, recruited occipital lobe areas, while verbs recruited parietal or motor cortex areas. In contrast, Longe et al. (2007, p. 1812) used functional magnetic resonance imaging (fMRI) to suggest that any differences in processing are due to “interactions between the morphological structure of nouns and verbs and the processing implications of this” and not grammatical class processing per se. In both cases, however, stimulus words were controlled using CELEX database frequency values and stimuli presented in visual form, so although the differences between the two apparently contradictory positions may ultimately be attributable to differences between technologies (EEG versus
fMRI), it cannot be concluded, as both sets of authors did, that either set of results was due to spoken language processing; rather, by using stimulus words based on normative values related to literacy processing and by presenting those stimuli in visual form, their results may be literacy not speech related. At this point, any effects caused by differences in speech and literacy processing at a neural level remain unknown.

To this end, the purpose of the current study was to compare normative values obtained directly from a prospective experimental population with corresponding values sourced from a selection of existing databases. The 140 words contained within the corpus were originally chosen as experimental stimuli for an EEG investigation of grammatical class (nouns versus verbs) processing, following in the style of both Pulvermuller et al. (1999) and Longe et al. (2007). As a result, original inclusion/exclusion criteria were based firstly on grammatical class purity (that is, the word does not function as both a noun and a verb in contemporary usage), secondly on stress typicality (typical versus atypical) and finally on syllable number (<3). Given that a number of linguistic features have been suggested as influencing neural responses (for discussion, see Pulvermuller et al., 1999), normative data were firstly gathered on the features of familiarity/frequency, concreteness and imageability from the prospective EEG experimental population. This was conducted via six, separate internet based surveys, three of which presented each stimulus word in written form, the remaining three in spoken form. Individual participant ratings for each stimulus word were then averaged across each sample to provide normative data values for each of the stimulus words. These average values were then compared with word form values sourced from three of the existing database sources; specifically CELEX, the BNC and the MRC. CELEX was chosen to follow in the design of Pulvermuller et al. (1999) and Longe et al. (2007). The BNC was chosen based on spoken word values displaying the best ecological validity during collection (that is, natural versus artificially natural speech). The MRC was chosen as it is the only database offering both concreteness and imageability normative values.

Based on the assumption that the use of these databases for sourcing normative linguistic feature values is acceptable practice for speech comprehension/perception paradigms, the following four hypotheses
were proposed. Firstly, to confirm that frequency (production) values are interchangeable with familiarity (comprehension) values, there should be a strong positive correlation between all four value sources across the corpus; that is, familiarity values obtained by the current study’s surveys (both written and spoken) should show a strong positive correlation with the frequency values sourced from CELEX and the BNC, and the familiarity values sourced from the MRC.

Secondly, to confirm that mode of presentation (written versus spoken) has no influence on frequency/familiarity values, it was expected that there would be no significant mean differences between BNC spoken frequency values and BNC written frequency values. Thirdly, to also confirm that mode of presentation (written versus spoken) has no influence on familiarity, concreteness or imageability values, it was predicted that there would be no significant mean differences within the current study’s corpus related to presentation type. Finally, if a database’s age does not distort normative values, then there should be no significant differences between the values obtained in the current study’s surveys and those sourced from the MRC (which, as the oldest of the databases is most likely to show an effect).

**Materials and Method**

**Participants**
Six separate, independent samples were recruited from the student and staff populations of the University of New England (UNE), Armidale, New South Wales, Australia and the wider Armidale and Northern Tablelands community. Participants were invited to click on one of six different URLs to access their survey of choice; specifically, Familiarity Written (FW), Familiarity Spoken (FS), Concreteness Written (CW), Concreteness Spoken (CS), Imageability Written (IW) and Imageability Spoken (IW). Each survey carried full human research ethics approval (HE10/26; HE11/045) while first year psychology students were eligible for course credit in exchange for participation. The median age range, across both males and females, for all of the separate surveys was 25-35 years. Further demographic details for each survey can be found in Table 1.
Bernadine Cocks, Jana Smith, and Graham A. Jamieson

Materials

Stimulus words. One hundred disyllabic words (50 x verbs, 50 x nouns) were pseudo-randomly selected from The Australian Concise English Dictionary 4th Edition (Moore, 2004). Of these, half of each grammatical class (25) displayed typical stress (that is, trochaic stress for nouns, such as “OB-ject”, and iambic stress for verbs, such as “ob-JECT”) with the remainder displaying atypical stress. Another 40 monosyllabic words were also randomly selected, again equally divided between nouns and verbs. This thus resulted in a total of 140 stimulus words. Although some of these words could function as another grammatical class type (e.g., adjective) none, according to the dictionary used, could be used interchangeably within the classes controlled for; that is, none of the nouns were accepted as being able to be used as verbs, and vice-versa. Grammatical class was checked using the dictionary entry, typicality was assigned by the researcher. See Appendix A for the complete stimulus word list.

Surveys. Stimulus words were presented to participants via one of six (FW, FS, CW, CS, IW or IS) online surveys hosted by Qualtrics (Qualtrics Labs Inc., Provo, UT, 2009). Based on the methodology of Gilhooly and Logie (1980) as well as several other studies (e.g., Cortese & Fugett, 2004; Della Rosa, Catricala, Vigliocco, & Cappa, 2010; Izura, Hernandez-Munoz

<table>
<thead>
<tr>
<th></th>
<th>FW</th>
<th>FS</th>
<th>CW</th>
<th>CS</th>
<th>IW</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total n</td>
<td>278</td>
<td>107</td>
<td>225</td>
<td>101</td>
<td>262</td>
<td>100</td>
</tr>
<tr>
<td>Male n</td>
<td>51</td>
<td>22</td>
<td>47</td>
<td>25</td>
<td>52</td>
<td>23</td>
</tr>
<tr>
<td>Female n</td>
<td>225</td>
<td>85</td>
<td>163</td>
<td>75</td>
<td>206</td>
<td>77</td>
</tr>
<tr>
<td>Did not matriculate High School (%)</td>
<td>4.3</td>
<td>2.1</td>
<td>3.8</td>
<td>7.0</td>
<td>3.1</td>
<td>7.0</td>
</tr>
<tr>
<td>Matriculated High School (%)</td>
<td>21.9</td>
<td>34.3</td>
<td>28.1</td>
<td>26</td>
<td>32.8</td>
<td>34</td>
</tr>
<tr>
<td>TAFE or Trade Course (%)</td>
<td>30.9</td>
<td>20.4</td>
<td>26.7</td>
<td>27</td>
<td>23.7</td>
<td>22</td>
</tr>
<tr>
<td>Undergraduate University degree (%)</td>
<td>26.2</td>
<td>22.2</td>
<td>24.8</td>
<td>29</td>
<td>23.3</td>
<td>23</td>
</tr>
<tr>
<td>Post-Graduate University Degree (%)</td>
<td>12.9</td>
<td>14.8</td>
<td>12.4</td>
<td>7.0</td>
<td>12.2</td>
<td>11</td>
</tr>
<tr>
<td>Other (eg Nursing certificate) (%)</td>
<td>3.5</td>
<td>5.6</td>
<td>3.8</td>
<td>4.0</td>
<td>3.8</td>
<td>3.0</td>
</tr>
</tbody>
</table>
What should be the Place of the Database?

& Ellis, 2005; Paivio, Yullie, & Madigan, 1968; Stadthagen-Gonzalez & Davis, 2006; Strain, Patterson & Seidenberg, 1995) each separate survey was presented with a Likert style response scale for each linguistic feature. For CW or CS, there were seven response options ranging from “Totally abstract” to “Totally concrete”. Concrete was overtly defined in the instructions to participants displayed at the beginning of each page as “solid or tangible, something you could physically touch” while abstract was defined as “intangible, more a feeling or an idea than a “thing”. IW and IS also had seven response options ranging from “very difficult” to “very easy”, with imageability defined as “how easily you can imagine what the word represents; that is, how easy it is to visualise what the word means”. For FW and FS, participants had five response options ranging from “never heard of it” to “know this word very well”. Although this varies slightly from Gilhooly and Logie’s (1980) methodology (a 5 point versus 7 point scale) it was thought that asking participants to decide between “know this word well” and “know this word very well” was too fine grained a distinction.

In the case of spoken word presentation surveys (FS, CS and IS), each stimulus word was recorded by a professional male voice on a Rode NT 2 microphone through a Phonic MM1002 mixer. Sound files were then normalised using Cool Edit Pro V1.1. Written word presentation used a default Qualtrics font and template previously purchased by the UNE.

**MRC.** The MRC Psycholinguistic Database (Coltheart, 1981) was used to retrieve normative data values for concreteness, imageability and familiarity. These values are derived from an averaging of norms collected from three databases; the Paivio, Yuille and Madigan norms (1968), the Toglia and Battig norms (1978), and the Gilhooley and Logie norms (1980), with each presented as a numeric value between 100 and 700 which were originally derived from a 7 point Likert scale (Wilson, 1988). As the MRC does not offer normative values for all three linguistic features for every word contained in its corpus, it was not possible to obtain appropriate values for all stimulus words used in the current study; thus, comparisons can only be made on a subset of words – concreteness (45), familiarity (51) and imageability (49).
Bernadine Cocks, Jana Smith, and Graham A. Jamieson

**BNC.** The BNC does not provide normative data for either concreteness or imageability. It does, however, offer word frequency scores per million for both written and spoken words. Thus it was possible to retrieve normative frequency data for 126 of the 140 spoken words in the current study, as well as 137 of the 140 written words. It should also be noted that the presentation of frequency values in the BNC makes direct comparison with either the MRC or current study’s surveys impossible although strong correlations between these different sources would suggest that they are all measuring the same underlying construct.

**CELEX.** The CELEX linguistic database was used to access written frequency values. Last updated in 1995, the written frequency values contained in CELEX were derived from texts composed by northern hemisphere (as opposed to Australasian) authors. Like the BNC, CELEX does not offer values for concreteness or imageability, nor does it offer frequency values for every word in the English language, therefore only written frequency values for 73 words could be obtained. Also, like the BNC, the presentation of the frequency values in the CELEX database makes direct comparison with either the MRC or current study’s surveys impossible.

**Procedure**

**Written surveys (FW, CW, IW).** Participants were invited to click on their choice of one of three separate URLs. Each URL lead to a different survey thus resulting in three separate, individual samples. The first page on each survey provided background and ethically required information including a statement regarding informed consent. Participants then progressed through their chosen survey by reading each word and clicking an on-screen button to respond on the Likert scale.

**Spoken (FS, CS, IS).** Similar to the written surveys, participants were invited to click on their choice of one of three separate URLs. Again, the first page on each survey provided background and ethical information including a statement on informed consent. Participants then progressed
through their chosen survey by clicking on the ‘play’ icon for each question to listen to the word, and then selected their response on the Likert scale.

Statistical analyses

Data from participants for each of the six separate surveys were initially screened for missing values. Individual participants who displayed >10% missing responses were excluded from subsequent analyses. Where <10% of the responses were missing (< 2% of the total sample), these were replaced by the mean value of the remaining scores for each item (word); that is, the mean of all participants’ responses for that word was used.

Data was then checked for assumption violations. As all surveys, the BNC, MRC and CELEX were significantly skewed, both square root and log 10 transformations were applied. As neither corrected the normality assumption violations, untransformed data were subsequently used for all analyses. It should be noted, however, that these violations of normality was not especially surprising given the nature of the surveys. It should also be noted that all surveys were skewed in the same direction.

Participant data was then averaged across individual words to allow item analysis based on mean ratings of each word in each condition. To match the values published by the MRC, these values were then multiplied by 100, as per the methodology employed by Gilhooley and Logie (1980).

Given the normality assumption violations and obviously differing sample sizes, only non-parametric tests were considered reliable and will be reported here. Firstly, Spearman’s Rank Order correlations were performed between the different sources (BNC, MRC, CELEX, current survey) on the familiarity/frequency dimension to test hypothesis 1. To test hypotheses 2 and 3, Wilcoxon signed rank tests were then used to compare spoken presentation word values with written word presentation values within both the BNC and current surveys. Finally, to test hypothesis 4, a Friedman two-way ANOVA, was used to examine any differences between the results of the current study’s surveys (both spoken and written types) and the MRC, across the three target linguistic features (familiarity, concreteness, imageability). All tests were conducted using PAWS V.18 (IBM, Chicago, IL), at a significance level of α=.05, two-tailed.
Results

Spearman’s correlations were performed between the written surveys (FW, CW, IW), the spoken surveys (FS, CS, IS) and linguistic database values. This correlational analysis allowed for confirmation that the current study’s surveys were measuring the same, or at least similar, underlying constructs and specifically for testing whether familiarity values in the current study’s surveys corresponded to frequency values sourced from the selected databases (hypothesis 1). Although a lack of corresponding data in the MRC reduced the sample size to \( n=51 \) (familiarity), \( n=45 \) (concreteness) and \( n=49 \) (imageability), a number of significant correlations were found. Of note were the near perfect correlations between the current study’s written and spoken survey type values irrespective of construct (that is, familiarity, concreteness and imageability). Also, of note were the strong positive correlations between the current study’s familiarity values (written), the MRC familiarity values, and both the BNC written and spoken values, although there was no significant relationship found with CELEX frequency values. Similarly, the current survey’s familiarity values

Table 2. Spearman’s Rho Rank Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>FW</th>
<th>CW</th>
<th>IW</th>
<th>FS</th>
<th>CS</th>
<th>IS</th>
<th>MRCF</th>
<th>MRCC</th>
<th>MRCI</th>
<th>BNCSF</th>
<th>BNCWF</th>
<th>CELEXF</th>
</tr>
</thead>
<tbody>
<tr>
<td>FW</td>
<td>1</td>
<td>.08</td>
<td>.45**</td>
<td>.94**</td>
<td>.12</td>
<td>.39**</td>
<td>.77**</td>
<td>.10</td>
<td>.13</td>
<td>.80**</td>
<td>.71**</td>
<td>.14</td>
</tr>
<tr>
<td>CW</td>
<td>1</td>
<td>.73**</td>
<td>.06</td>
<td>.97**</td>
<td>.82**</td>
<td>.07</td>
<td>.79**</td>
<td>.83**</td>
<td>-.15</td>
<td>-.07</td>
<td>-.07</td>
<td></td>
</tr>
<tr>
<td>IW</td>
<td>1</td>
<td>.42**</td>
<td>.74**</td>
<td>.92**</td>
<td>.15</td>
<td>.53**</td>
<td>.61**</td>
<td>.09</td>
<td>.07</td>
<td>-.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FS</td>
<td>1</td>
<td>.10</td>
<td>.37**</td>
<td>.72**</td>
<td>.03</td>
<td>.07</td>
<td>.75**</td>
<td>.68**</td>
<td>.18*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td></td>
<td></td>
<td></td>
<td>.84**</td>
<td>.07</td>
<td>.77**</td>
<td>.82**</td>
<td>-.09</td>
<td>-.04</td>
<td>-.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1</td>
<td></td>
<td></td>
<td>1.9</td>
<td>.70**</td>
<td>.76**</td>
<td>.06</td>
<td>.08</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRCF</td>
<td>1</td>
<td></td>
<td></td>
<td>.78**</td>
<td>.03</td>
<td>.07</td>
<td>.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRCC</td>
<td>1</td>
<td></td>
<td></td>
<td>.76**</td>
<td>.06</td>
<td>.08</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRCI</td>
<td>1</td>
<td></td>
<td></td>
<td>.84**</td>
<td>.18*</td>
<td>.81**</td>
<td>.62**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**p<.01, *p<.05; MRCF= MRC Familiarity, MRCC= MRC Concreteness, MRCI= MRC Imageability, BNCSF = BNC spoken frequency, BNCWF = BNC written frequency, CELEXF = CELEX frequency**
What should be the Place of the Database?

(spoken) showed a strong positive correlation with the MRC familiarity values, and both the BNC written and spoken values. There was also a significant, although smaller correlation with CELEX frequency values. Additionally, the current study’s spoken and written surveys were near perfectly correlated on the dimensions of imageability and concreteness, as well as being strongly correlated with the MRC values for concreteness and imageability, irrespective of presentation type (written versus spoken). Finally, strong positive correlations were also found between CELEX sourced frequency values and familiarity values sourced from the MRC and the BNC (both written and spoken). See Table 2 for the complete Spearman’s correlation matrix.

**Word Stimuli**

To test whether frequency/familiarity values would vary significantly between type of presentation (written versus spoken) across the corpus (hypothesis 2), a Wilcoxon signed rank test was performed between the BNC written and spoken frequency values. This showed that written presentation frequency values were, on average, higher than spoken presentation frequency values, \( T=149, z=-9.42 \) (corrected for ties), \( N- Ties=127, p<.001, \) two-tailed, with a large effect size \( r=.84 \).

Wilcoxon signed rank tests were then performed between the current study’s written and spoken surveys, across all three linguistic features (familiarity/frequency, concreteness, imageability) to test hypothesis 3. In keeping with the differences found within the BNC corpus, a significant difference was found between presentation style for familiarity, \( T=258.50, z=-9.34 \) (corrected for ties), \( N- Ties=131, p<.001, \) two-tailed, with a large effect size \( r=.82 \), as well as (2) concreteness, \( T=408.50, z=-9.23 \) (corrected for ties), \( N- Ties=126, p<.001, \) two-tailed, with a large effect size \( r=.82 \), and (3) imageability, \( T=144.50, z=-9.68 \) (corrected for ties), \( N- Ties=133, p<.001, \) two-tailed, with a large effect size \( r=.84 \).

To finally examine if there were any significant differences between the current survey’s results and the MRC, and thereby test hypothesis 4, a Friedman two-way ANOVA was then performed. This showed significant variation between each presentation type (spoken and written) and the MRC across each linguistic feature. Specifically, significant results were
found for familiarity, $\chi^2_F = 43.17$ (corrected for ties), $df=2$, $N$-Ties=48, $p < .001$; concreteness, $\chi^2_F = 43.03$ (corrected for ties), $df=2$, $N$-Ties=43, $p < .001$; and imageability, $\chi^2_F = 55.62$ (corrected for ties), $df=2$, $N$-Ties=44, $p < .001$.

**Discussion**

Partially confirming predictions, strong positive associations were found between the current study’s surveys and both the MRC and the BNC on the familiarity/frequency dimension. This was taken as being sufficient evidence that the current study’s familiarity values were measuring the same underlying construct as the other databases’ (BNC and MRC) frequency/familiarity values. Contrary to predictions, however, there was no significant correlation between CELEX normative frequency values and the current study’s familiarity values, although CELEX did display small to moderate positive correlations with both the MRC familiarity values and BNC written frequency values. The exact reasons for this lack of relationship are unclear, although it thought likely a reflection of differences in value gathering; that is, the CELEX values used were derived from the frequency of words appearing in written text or environments whereas the current study provided a more naturalistic measure of spoken language use. Irrespective of this, the results do suggest that CELEX written frequency values (production) are not interchangeable with familiarity (comprehension) values.

Also contrary to predictions, significant differences were found based on presentation type (written versus spoken) for both the BNC and the current study across all linguistic features; that is, stimuli presented in written form were consistently rated significantly higher than those presented in auditory form, irrespective of whether the construct under consideration was familiarity/frequency, concreteness of imageability. Again the exact reasons are unclear within the current study, however, it may be related to temporal influences on human language perception. In the case of written stimuli, all visual features are available for processing from the very first millisecond the stimulus becomes visible. With auditory stimuli, however, stimulus features unfold over time (at least 500 ms for each of the current study’s
What should be the Place of the Database?

stimuli). This difference in temporal *availability* of features for processing may thus influence subjective feelings of familiarity; by having longer to process the complete range of stimulus features, the participants might, at a pre-conscious level, be influenced to rate written stimuli higher on this basis alone. In a similar vein, *seeing* words may artificially inflate concreteness and imageability values simply because print appears more concrete and imageable than sound. Although purely conjectural at this point, it is an area of human language processing that appears to have garnered little (if any) empirical interest.

Finally, and again contrary to predictions, a database’s age does appear to distort normative values with the oldest of the databases examined (the MRC) showing significant mean differences compared to the current study across all linguistic features under investigation. While this may be a reflection of the actual sample corpus used or subtle dialectal differences between the different value sources, it is considered more likely that it is a reflection of the age of this database. As demonstrated by Senghas and Coppolla (2001), language can evolve substantially in as little as twenty years. The MRC, however, is derived from data as old as 50 years. Thus, despite the MRC’s ease of use and ability to provide normative data across a range of linguistic features, it appears somewhat past its use-by-date. This is, perhaps, concerning given that a Google Scholar search (conducted 21/11/13) restricted to only 2013 publications returned in excess of 250 papers citing the use of MRC values.

As with any research the current study is not without limitations. Firstly, women outnumbered men by three to one in all six of the current study’s surveys. As a result, gender differences in language use may have skewed the current study’s findings. Although this may impact on the interpretation of the comparisons between the current study’s surveys and the databases used, it does not necessarily have any bearing on the significant differences found between presentation types, especially as the BNC showed a similar significant difference between spoken and written values. Further research into gender differences influencing the perception of familiarity, concreteness and imageability may thus be warranted.

A second limitation of the current study was that first year university students studying psychology represented as many as 90% of participants
in individual surveys, thus the values obtained in the current study cannot be generalised to the broader population. It should be noted, however, that psychology students are a prime source of participants for language research paradigms both within Australasia and overseas. As a result, the values obtained here may be more representative of the ‘populations’ used in language research than those published by databases and, as such, a more valid measure for controlling potential confounding variables.

Finally, it became apparent post hoc that two of the stimulus words used may have switched grammatical class in the conversion from written to spoken presentation. Specifically, the words ‘baste’ and ‘pact’, designated as nouns in the written surveys, could easily have been perceived in the spoken surveys as their homonyms, ‘based’ and ‘packed’, both verbs. Given, however, that these words represent less than 1.5% of the overall data, it seems unlikely that any impact on results would be of any great consequence. Paradoxically, it does highlight how easy it is to overlook the slight differences between written and spoken language, even by language researchers who are critiquing language research! Despite such limitations, the results of the current study highlight that normative database values should be used with a degree of caution in language research paradigms, especially those involving speech perception/comprehension.

References

What should be the Place of the Database?

Experimental Psychology, 33A, 497-505


*Manuscript received: Sep 05, 2013, in revised form: Dec 09, 2013*
Appendix A: Complete stimulus word list.

Absence  Act  Addle  Affair
Applause  Argue  Art  Attract
Badge  Baggage  Baste  Become
Believe  Brigade  Burgle  Canal
Cape  Carve  Cater  Clinic
Consume  Dairy  Dame  Deem
Degree  Destroy  Differ  Domain
Eat  Effort  Elate  Enter
Erode  Estate  Event  Fabric
Façade  Fact  Fend  Follow
Fondle  Forbid  Gadget  Gobble
Govern  Grenade  Grow  Guitar
Hammock  Harass  Held  Hinder
Hotel  Ignite  Imp  Indulge
Infant  Intent  Journal  Jungle
Keg  Kept  Kidnap  Lagoon
Lake  Languish  Learn  Legend
Locate  Maelstrom  Manage  Manure
Meal  Merge  Monsoon  Mutate
Napkin  Narrate  Nourish  Nymph
Oath  Obey  Ocean  Offence
Oppose  Ordeal  Oust  Pact
Pageant  Pave  Percent  Perform
Pierce  Prevent  Punish  Quart
Quartet  Quash  Quota  Ransack
Realm  Reap  Recede  Refer
Relief  Rodent  Sage  Saloon
Sarong  Seduce  Seek  Sever
Spasm  Suckle  Suppress  Tamper
Tarry  Teach  Terrain  Tissue
Twerp  Udder  Undo  Usurp
Valance  Valve  Vanquish  Vend
Venom  Vibrate  Wallet  Warn
Wither  Withstand  Wreath  Yearn
Yeast  Yodel  Yogurt  Zeal
Zombie