Acta Psychologica 150 (2014) 80–84

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Norms of age of acquisition and concreteness for 30,000 Dutch words

Marc Brysbaert [a,*], Michaël Stevens [a], Simon De Deyne [b], Wouter Voorspoels [b], Gert Storms [b]

[a] Ghent University, Belgium
[b] University of Leuven, Belgium

ARTICLE INFO

Article history:
Received 11 March 2014
Received in revised form 22 April 2014
Accepted 23 April 2014
Available online 13 May 2014

PsycINFO:
2340
2720

Keywords:
Norms
Age of acquisition
Concreteness
Word recognition

ABSTRACT

Word processing studies increasingly make use of regression analyses based on large numbers of stimuli (the so-called megastudy approach) rather than experimental designs based on small factorial designs. This requires the availability of word features for many words. Following similar studies in English, we present and validate ratings of age of acquisition and concreteness for 30,000 Dutch words. These include nearly all lemmas language researchers are likely to be interested in. The ratings are freely available for research purposes.

1. Introduction

Research on word recognition is rapidly changing. Authors realise that the traditional small-scale factorial experiments are not the best approach because they lack power (Keuleers, Diependaele, & Brysbaert, 2010), do not give information about the full range of variables (Kuperman, Estes, Brysbaert, & Warriner, in press), and are open to experimenter bias in stimulus selection (Forster, 2000; Kuperman, in press). A better approach is to treat word recognition studies not as experiments in which word features can be manipulated but as correlational studies in which covariations between word features and word processing performance can be assessed (Baayen, Feldman, & Schreuder, 2006; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Lewis & Vladeanu, 2006). As a result, researchers have collected word processing times for thousands of words in so-called lexicon projects. Thus far, this happened in American English (Balota et al., 2007), Dutch (Keuleers et al., 2010), Malay (Yap, Rickard Liow, Jalil, & Faizal, 2010), French (Ferrand et al., 2010), British English (Keuleers, Lacey, Rastle, & Brysbaert, 2012), and Chinese (Sze, Rickard Liow, & Yap, in press).

At the same time, an optimal use of the lexicon projects requires information about the word features for (ideally) the entire database. This is easy for word variables that can be calculated on the basis of the words themselves or corpus analyses, such as word length, various measures of word frequency, and similarity to other words but requires a major investment for variables that are based on subjective ratings.[1] These are variables like age of acquisition, concreteness, imageability, familiarity, valence, and arousal. They are investigated for their own sake or must be controlled for in order not to confound the effect of the variable of interest.

The situation is rapidly improving for the English language, where age-of-acquisition ratings have been collected for 30,000 words (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), affective ratings for 14,000 words (Warriner, Kuperman, & Brysbaert, 2013), and concreteness ratings for 40,000 words (Brysbaert, Warriner, & Kuperman, in press). The main reason for this improvement is that in English, one can make use of Amazon Mechanical Turk, a service created by the company Amazon where Internet users can earn a small amount of money by doing so-called Human Intelligence Tasks. These are usually short rating or translation tasks. Because there are several tens of thousands of Mechanical Turk workers, large-scale rating studies can be done in a matter of weeks at an affordable price. In addition, if some basic controls are included, the ratings are as reliable and valid as those collected under traditional laboratory circumstances (for evidence, see the references above).

* Corresponding author at: Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Gent, Belgium. Tel.: +32 9 264 94 25; fax: +32 9 264 64 96.
E-mail address: marc.brysbaert@ugent.be (M. Brysbaert).

[1] An extra complication is that it is difficult to secure funding for the collection of such ratings, because research councils seem to have an aversion for research proposals that are not driven by theory falsification, even though good hypothesis testing critically depends on access to this information.

http://dx.doi.org/10.1016/j.actpsy.2014.04.010
0001-6918/

The situation is different for other languages because Amazon Mechanical Turk is based in the United States and has much fewer users/workers in languages other than English or Spanish (also the payment happens in dollars via the American branch of the company). This means that Amazon Mechanical Turk is a less interesting tool to collect data for languages such as Dutch. However, Moors et al. (2013) recently proposed an alternative solution. They showed that asking a limited group of participants to rate a list of 4,300 words returns the same outcome as the traditional approach of asking a large number of participants to rate 300 words each. The costs for paying the participants are the same, but the logistics become much more feasible. Also, participants are more interested and motivated when they can earn more money (because of the larger time investment).[2]

Arguably, the two most important word norms based on subjective ratings are age of acquisition (AoA) and concreteness. AoA refers to the age at which a word has been acquired and explains some 5% of variance in lexical decision times after the effects of word frequency, word length, and similarity to other words have been partialed out (Kuperman et al., 2012). This is even more when a suboptimal word frequency measure is used (Brysbaert & Cortese, 2011). The impact of AoA is due to the fact that the order of acquisition is an important variable in the organisation of the mental lexicon and the semantic system (Bai, Ma, Dunlap, & Chen, 2013; Catling, Dent, Preece, & Johnston, 2013; Cortese & Schock, 2013; Cuetos, Herrera, & Ellis, 2010; Palmer & Havelka, 2010) and to the fact that AoA is an important proxy for estimating the cumulative frequency with which people have come across words in their life (Lete & Bonin, 2013).

Concreteness evaluates the degree to which a concept denoted by a word refers to a perceptible entity. It is an important variable in memory research ever since Paivio formulated his dual-coding theory (Paivio, 1971, 2013). According to this theory, concrete words are easier to remember than abstract words because they activate perceptual memory codes in addition to verbal codes. The variable gained extra interest within the embodied view of cognition (Barsalou, 1999; Fischer & Zwaan, 2008; Wilson, 2002), certainly after it was established that words referring to easily perceptible entities co-activate the brain regions involved in the perception of those entities, and that action-related words co-activate the motor cortex involved in executing the actions. On the basis of these findings, Vigliocco, Vinson, Lewis, and Garrett (2004) (see also Andrews, Vigliocco, & Vinson, 2009) presented a semantic theory, according to which the meaning of concepts depends on experiential and language-based connotations to different degrees. Some words are mainly learned on the basis of direct experiences; others are mostly used in text and discourse.

Concreteness is also much researched in psycholinguistics. These are a few examples of recently examined topics related to concreteness. Are there hemispheric differences in the processing of concrete and abstract words (Oliveira, Perea, Ladera, & Gamito, 2013)? Does concreteness affect bilingual and monolingual word processing (Barber, Otten, Kousta, & Vigliocco, 2013; Connell & Lynott, 2012; Gianico-Relyea & Altarriba, 2012; Kaushanskaya & Rechtzigel, 2012)? Do concrete and abstract words differ in affective connotation (Ferré, Guasch, Moldovan, & Sánchez-Casas, 2012; Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011)? Do neuropsychological patients differ in the comprehension of concrete and abstract words (Loiselle et al., 2012)?

Imageability and familiarity are less interesting variables because imageability is highly correlated with concreteness and seems to stress the visual modality too much (Connell & Lynott, 2012). The importance of familiarity is likely to be minimal, once one has a good word frequency measure and information about AoA (Brysbaert & Cortese, 2011). Valence and arousal have recently gained interest (e.g., Kuperman et al., in press) but could not be included in the present study (see, however, Moors et al., 2013, who collected values for 4,300 words).

AoA ratings were available for a few thousand words in Dutch. Ghyselinck, De Moor, and Brysbaert (2000) collected norms for some 3,000 short words. Ghyselinck, Custers, and Brysbaert (2003) collected ratings for a further 2,300 words from much used semantic categories (such as clothes, animals, utensils, birds, etc.). Finally, Moors et al. (2013) collected ratings for 4,300 words. To our knowledge, there are no large collections of concreteness ratings, but imageability norms were collected by Van Loon-Vervoorn (1985) for 6,100 words. The correlations between concreteness and imageability reported in the literature range from 0.78 to 0.85 (Friendly, Franklin, Hoffmann, & Rubin, 1982; Gilhooly & Logie, 1980; Paivio, Yuille, & Madigan, 1968).

Below, we describe the collection of concreteness and AoA ratings for 30,000 Dutch words.

2. Method

2.1. Stimulus materials

On the basis of dictionaries and corpus analyses, we selected a list of 30,000 'interesting' words. Interesting was defined in terms of the following:

1. Words are lemmas (unless the inflected form is highly frequent; e.g., 'eyes' in addition to 'eye').
2. No proper nouns are used.
3. The words are likely to be known to the participants.
4. No long, transparent compound words are included. Dutch is a language in which compounds do not have spaces, meaning that hundreds of thousands of words can be made by combining base words. Bertram and Hyona (2003) reviewed the reasons why low frequency, long, and transparent compounds are unlikely to be represented in the mental lexicon (they are parsed into their constituent meanings on the spot).

2.2. Participants

The participants were 74 students and scientific collaborators from Ghent University who completed the AoA lists and 75 students and collaborators from the University of Leuven who completed the concreteness lists. Of the Ghent participants, 11 were male and 63 female. Their mean age was 21.8 years (range, 18–32 years). Of the Leuven participants, 21 were males and 54 females. Their average age was 25.08 years (range, 17–63 years). Ghent and Leuven are two towns in Flanders (the Dutch speaking half of Belgium) separated by 80 km. Moors et al. (2013) found no differences in the ratings of the two universities for the variables they investigated, and there are no reasons to expect this would be otherwise for the present ratings. Still, as a precaution, the Ghent students provided the AoA norms, given that all previous AoA ratings in Dutch were collected there. The Leuven students provided the concreteness norms. More students started the study, but they are not included in the analyses because they did not return their list, arguably because they lost interest after a few trials.

2.3. Methods

For the concreteness ratings, the master stimulus list was divided into five lists of 6,000 words each. Each participant got a different permutation (15 raters per list). The lists started with the same 10 calibrator words covering the entire range of values from very concrete to very abstract, based on the authors' judgment and the imageability ratings of Van Loon-Vervoorn (1985). The concreteness instructions

[2] Language researchers seem to have a peculiar aversion to time intensive studies. The most often mentioned reasons are fatigue effects and lack of motivation (as if people are not used to working for a few hours at a task). Other objections are practice effects and long term priming (as if participants in psychology experiments should be uncertain about the task they are doing). As it happens, there is good evidence that you get better data if the participants have some experience with the task. Certainly for the lexicon projects it is becoming clear that time intensive studies with a limited group of participants are providing less noise than short studies with a large group (Keuleers et al., 2010, 2012).