Behav Res (2012) 44:998–1006
DOI 10.3758/s13428-012-0195-z

Checking and bootstrapping lexical norms by means of word similarity indexes

Yves Bestgen · Nadja Vincze

Published online: 7 March 2012

Abstract In psychology, lexical norms related to the semantic properties of words, such as concreteness and valence, are important research resources. Collecting such norms by asking judges to rate the words is very time consuming, which strongly limits the number of words that compose them. In the present article, we present a technique for estimating lexical norms based on the latent semantic analysis of a corpus. The analyses conducted emphasize the technique's effectiveness for several semantic dimensions. In addition to the extension of norms, this technique can be used to check human ratings to identify words for which the rating is very different from the corpus-based estimate.

Keywords Lexical norms · Automatic estimation · Latent semantic analysis (LSA) · Valence · Arousal · Dominance · Concreteness · Imagery

Electronic supplementary material The online version of this article (doi:10.3758/s13428-012-0195-z) contains supplementary material, which is available to authorized users.

The Supplemental material is a compressed archive file which is associated with this manuscript: http://www.psor.ucl.ac.be/personal/yb/Doc/BestgenVincze.zip. It contains five txt files. The file names indicate the norms: valence, arousal, dominance, concreteness and imagery. In every individual file, each line provides certain elements in the following order, separated by a space: the word, the value estimated by DIC-LSA, and the value according to the norms when available. When this value is not available, it is replaced by a period.

Y. Bestgen · N. Vincze
Université Catholique de Louvain,
Louvain-la-Neuve, Belgium

Y. Bestgen (✉)
Psychology Department, Université Catholique de Louvain,
10 Place du Cardinal Mercier,
1348 Louvain-la-Neuve, Belgium
e-mail: yves.bestgen@psp.ucl.ac.be

For more than forty years, many psychological studies have been conducted in a variety of languages to collect norms. Such studies have focused on formal and semantic properties of words, such as frequency of use, age of acquisition, familiarity, concreteness, imagery, and emotional valence (for indices of these norms, see, e.g., Bradshaw, 1984; Proctor & Vu, 1999). These norms mainly serve for selecting experimental materials used in studies regarding the relationship between imagery, valence, or familiarity and ease with which a person is able to understand a word (e.g., Jessen et al., 2000; Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011; Kroll & Merves, 1986). When the size of the norms is large enough, they are also used in regression analysis to predict the reaction times (RTs) obtained in word recognition experiments (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Brysbaert & Cortese, 2011). Such norms are also employed to develop reading tests (Desrochers & Saint-Aubin, 2008), to analyze texts written by foreign language learners (Dewaele & Pavlenko, 2002), and to measure emotional expression in language (Bestgen, 1994; Cohen, Minor, Najolia, & Hong, 2009; Pennebaker, Mehl, & Niederhoffer, 2003).

Some of these properties are gathered through automatic counting procedures applied to lexical databases or to corpora (e.g., oral and written word frequencies, orthographic or phonological neighborhoods, number of homophones). Other properties, such as familiarity, subjective frequency, concreteness, valence, arousal, and dominance are collected by asking participants to rate the words according to these dimensions. Collecting such norms is very time consuming, which strongly limits the number of words that compose them. Being able to automatically extend the norms would greatly facilitate their construction and offer new perspectives for optimizing word selection in factorial experiments and for drawing large samples for multiple-regression studies. Another problem

with these empirical norms is that ratings regarding a specific dimension can be contaminated by other word properties. This can be observed when a rater determines that an unfamiliar word is harder to picture than a word with which they are more familiar (Desrochers & Thompson, 2009). When the same dimension can be measured by counting procedures and by human ratings, as is the case with word frequency, a dual approach is recommended "in a cross-validation strategy in order to compensate for their respective weaknesses" (Desrochers & Thompson, 2009, p. 547). Currently, however, dimensions collected both through rating studies and by counting procedures applied to corpora are very rare.

The goal of the present article is to propose a technique to implement this type of estimate from corpora. This technique would serve not only to compare these estimates with human ratings, but also to extend the norms without having to call upon new raters. In the following section, we present work in computational linguistics that has established a foundation for such approaches. Then, we present a series of tests that were performed to evaluate the effectiveness of the technique for estimating different dimensions.

Estimation of lexical norms by automatic procedures

In recent years, researchers in the field of computational linguistics have become interested in some of the norms collected in psychology because of their usefulness for opinion mining—a relatively new area of research that aims to categorize texts according to their expressed sentiments (Pang & Lee, 2008). Most of the proposed approaches require emotional lexicons (Valitutti, Strapparava, & Stock, 2004). These lexicons contain words tagged with their affective valence (also referred to as *affective polarity* or *semantic orientation*) that indicates how strongly a word conveys a positive or a negative connotation. Because these lexicons must be the broadest possible, automatic techniques have been proposed to construct them.

Among these techniques, one can distinguish two types of approaches: those based on linguistic resources such as WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) and those based on corpora. The approaches based on linguistic resources generally estimate the similarity between the words from their relation of synonymy (e.g., Esuli & Sebastiani, 2006; Kamps & Marx, 2002; Kim & Hovy, 2004). Starting from a small set of words whose valence is known, a bootstrapping algorithm is run through the synonymic and antonymic links, and assigns the same orientation to the synonymous words and the reverse orientation to the antonymic words. Kamps and Marx were most likely the first to propose such a procedure by deriving a graph from WordNet in which each node represents a word, and edges connect any pairs of synonymous words. This graph is used to assign values to nodes according to the minimal path lengths to the adjective *good* and to the adjective *bad*. The principal limitation of this technique is that it applies only to adjectives connected by synonymous relations to the two seed words. Esuli and Sebastiani (2006) extended this approach by developing SentiWordNet, a procedure that assigns to each WordNet synset a positive and a negative value by means of a semisupervised learning procedure.

Approaches that rely on corpora lack information regarding synonymy and thus have to estimate semantic similarities differently. Hatzivassiloglou and McKeown (1997) have proposed an algorithm to infer the semantic orientation of adjectives based on an analysis of their co-occurrences with conjunctions. Turney and Littman (2002, 2003) and Bestgen (2002, 2008) developed more general techniques since they allow estimation of the valence of any term found in a corpus on the basis of its semantic proximity to other words whose valence is known. To determine the semantic proximity between two words, these researchers relied on latent semantic analysis (LSA), a mathematical technique for extracting a "semantic space" from large text corpora based on the statistical analysis of the set of co-occurrences in a corpus (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer, Foltz, & Laham, 1998). This technique, which is just one among many methods available to estimate semantic proximity (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Lund & Burgess, 1996), has often been used in cognitive and educational psychology to find words or texts that are most similar to a target item (Landauer, McNamara, Simon, & Kintsch, Landauer et al. 2007). The starting point of the analysis is a lexical table that contains the frequencies of every word in each of the text segments included in the corpus. This table is submitted to a singular value decomposition, which extracts the most significant orthogonal dimensions. In this semantic space, the meaning of a word is represented by a vector, and the semantic proximity between two words is estimated by the cosine between their corresponding vectors.[1]

In a technique called SO–LSA (semantic orientation based on latent semantic analysis), Turney and Littman (2003) used LSA to estimate the semantic proximity between a target word and 14 benchmarks: seven positive benchmarks (*good*, *nice*, *excellent*…) and seven negative benchmarks (*bad*, *nasty*, *poor*…). A word is considered as positive if it is closer to the positive benchmarks and further away from the negative benchmarks. Turney and Littman (2003) evaluated the effectiveness of their technique by comparing the predicted orientation of words with those defined in the General Inquirer

[1] The word similarities may also be calculated without the use of LSA, but, in this case, very large corpora are necessary, such as all the English texts available on the Internet, that is to say, some 100 billion words as in Turney and Littman (2003).

Lexicon (Stone, Dunphy, Smith, & Ogilvie, 1966), which contains a list of 3,596 English words with positive or negative labels. Out of a corpus of 10 million words, SO-LSA labeled 65% of the words correctly.

The principal difference between DIC–LSA—the technique proposed by Bestgen (2002)—and SO-LSA relies on the benchmarks used to evaluate a word. Whereas SO–LSA uses a few benchmarks selected a priori, DIC–LSA is based on dictionaries (the name used in content analysis to refer to norms) that contain several hundred words rated by judges on the pleasant–unpleasant scale (Bestgen, 1994; Bradley & Lang, 1999; Heise, 1965). To determine the emotional valence of a word according to its co-occurrence with other words in a corpus, a specific set of benchmarks is selected from the dictionary. More precisely, the unknown valence of a word corresponds to the average valence of its 30 nearest neighbors, the neighborhood being identified on the basis of the cosine in the semantic space.[2] To evaluate this technique, Bestgen (2002) compared the estimated values for French words with their actual values according to the dictionary and obtained correlations ranging from .56 to .70.

The main limitation of all the aforementioned techniques, excluding DIC–LSA, lies in the need to provide them with a handful of manually chosen seed words. To our knowledge, only Kamps and Marx (2002) proposed a pair of seed words for dimensions other than valence such as activity (active and passive) and potency (strong and weak). It should be noted, however, that their technique functions only with adjectives and that its effectiveness regarding activity and potency has not been evaluated. DIC–LSA can theoretically be applied to any dimension in any language, provided that the dimension can be estimated on the basis of semantic similarity. This is to say that the value of a word on a dimension can be estimated from its semantic similarity to other words whose value is already known. Checking the validity of this assertion was the objective of the study reported below. Five sets of semantic norms from two different studies were selected: valence, arousal, and dominance reported in the ANEW norms (Emotional Norms for English Words; Bradley & Lang, 1999), and concreteness and imagery from Gilhooly and Logie (1980). These dimensions have been widely studied for their effects on word processing, although heightened interest in arousal and dominance is a more recent development (Bayer, Sommer, & Schacht, 2010; Tipples, 2010). The semantic space was extracted from what is probably the most widely used corpus in psychological studies based on LSA: the TASA corpus (Landauer et al., 1998).

[2] A weighted average (by the cosine between each neighbor and the target word) can also be used, but experiments did not show any benefit of using this alternative formula.

Method

Lexical norms

The ANEW norms (Bradley & Lang, 1999) consist of 1,034 words rated by groups of eight to 25 judges. Their task was to indicate the emotional reaction evoked by specific words on three 9-point scales: valence (*negative, unpleasant* = 1; *positive, pleasant* = 9), arousal (*calm* = 1; *excited* = 9), and dominance (*feeling dominated* = 1; *feeling dominant* = 9). To assess these dimensions, Bradley and Lang utilized the Self-Assessment Manikin, an affective rating system based on nonverbal pictorial scales.

The concreteness and imagery norms for 1,944 words were collected by Gilhooly and Logie (1980). Concreteness ratings were provided by 35 participants on a 7-point scale that ranged from *concrete* (1) to *abstract* (7), whereas 37 participants provided imagery ratings on a 7-point scale ranging from words that failed to produce mental images or did so with difficulty (1) to words producing images readily (7).

Corpus for computing the word similarities

The semantic space used to compute word similarities was extracted from the General Reading up to 1st year college TASA corpus to which T.K. Landauer (Institute of Cognitive Science, University of Colorado, Boulder) provided access. The version utilized contains 44,486 documents and approximately 12 million words. This corpus was lemmatized by means of the TreeTagger (Schmid, 1994). In addition, a series of functional words (*and*, *be*, *the*, *that*…) were removed as well as all the words whose total frequency in the corpus was lower than 10. The resulting matrix of co-occurrences was submitted to a singular value decomposition, and the first 300 eigenvectors were retained.[3]

Results

The DIC–LSA technique as previously described was utilized to estimate the score on a dimension of any word included in the semantic space, disregarding whether it does or does not belong to the norms. The score of a word is equal to the average score of its *k* nearest neighbors whose score on that dimension is known (i.e., words that are

[3] Using the 300 first eigenvectors often produces the best results in LSA studies (Landauer, Laham, & Derr, 2004). In subsidiary analyses, this parameter was nevertheless varied per hundreds between 100 and 500. One hundred eigenvectors led to performances that were clearly below 300. Very little differences were observed for 200 to 500 eigenvectors. For some norms, results for 400 and 500 were very slightly better than those for 300, but the reverse was true for other norms

included in the norms). To evaluate the efficacy of this technique, only words present in the norms can be used. It is important to mention that a given word is never considered as one of its *k* nearest neighbors. It follows that the score this word received in the norms is never used to estimate it with DIC–LSA. The measures of efficacy reported below are thus obtained by a leave-one-out cross-validation technique often used in discriminant analysis, but also in regression analysis (Lachenbruch & Mickey, 1968; Stone, 1974).

The procedure estimated the score of the 17,350 terms present in semantic space according to the five norms. Among those words, 953 (of 1,034) are present in the ANEW norms and thus can be used to evaluate the efficacy of this technique. For the same reason, only 1,703 of the 1,944 words of Gilhooly and Logie's (1980) norms can be used in the following analyses.

Correlation between human ratings and automatic estimates

The efficacy of the automatic procedure to estimate the ratings on a semantic dimension was first assessed by means of the Pearson's correlation coefficient between the estimated and actual values. Table 1 shows these correlations for a series of values of the parameter *k* (the number of nearest neighbors included in the estimates). For comparison, the technique that Turney and Littman (2003) proposed for valence led to a correlation of .52 for this dimension, which is much lower than the values obtained by DIC–LSA for most *k*s.

For all of the dimensions, the correlations increase with larger *k*s until this parameter reaches 30. Thereafter, there is little change in the correlations. Large differences between semantic dimensions are observed; DIC–LSA proves to be more effective for concreteness than for imagery and valence, whereas arousal and dominance lead to the lowest correlations.

Table 1 Correlation between human ratings and DIC–LSA estimates for a range of values of the parameter k

K	Valence (*N*=951)	Arousal (*N*=951)	Dominance (*N*=951)	Concreteness (*N*=1,703)	Imagery (*N*=1,703)
1	.47	.32	.37	.65	.56
2	.55	.40	.44	.71	.61
3	.60	.46	.49	.73	.63
4	.64	.47	.51	.75	.65
5	.66	.48	.53	.76	.66
10	.68	.51	.56	.78	.68
20	.70	.55	.59	.79	.69
30	.71	.56	.60	.79	.70
40	.71	.56	.60	.79	.69
50	.71	.56	.60	.79	.69

These differences in efficacy between dimensions must be qualified by two additional observations. First, the lower efficacy for a certain dimension could find its origin when raters faced greater difficulty in using the corresponding scales. An analysis of the standard deviation of the ratings for each word, as was reported by Bradley and Lang (1999) indicates that the inter-rater variability is much larger for arousal (mean of the standard deviations=2.37, *SD*=0.31) than for dominance (*M*=2.06, *SD*=0.37), whereas valence results in the lowest inter-rater variability (*M*=1.65, *SD*=0.39). The differences between these three values are statistically significant (*t* test for paired data, all *p*s<.001). The dimensions for which DIC–LSA is the least effective also correspond to those with which the raters agreed less. A similar difference is observed for the imagery (*M*=1.61, *SD*=0.27) and concreteness (*M*=1.43, *SD*=0.35) ratings collected by Gilhooly and Logie (1980; *t* test for paired data, all *p*s<.001).

Second, imagery deserves a special analysis because of the strong relationship between imagery ratings and raters' familiarity with the words (Desrochers & Thompson, 2009). People have difficulty forming a mental image for words with which they are not particularly familiar. When one deliberately decides to employ a specific word, he or she is mostly likely familiar with the word. Therefore, a rare, but very imageable word, such as *yucca* or *phaeton* (see below), should occur in contexts containing other highly imageable words. This line of reasoning suggests that the correlation between imagery ratings and DIC–LSA estimates should be larger if words that are less familiar to the raters are removed from the data. This hypothesis can be easily tested because the Gilhooly and Logie's (1980) norms include familiarity ratings. These ratings were used to order words from the least familiar to the most familiar. Thereafter, the correlation between the estimated and actual imagery scores was computed in a repetitive way by removing each time the least familiar word remained in the data. The results of this analysis are presented in Fig. 1, in which the *x*-axis indicates the percentage of words that was removed from the data (from 0% to 75%) and the *y*-axis represents the correlation. As can be observed, the correlation strengthens as the percentage of removed data increases from 0% to 5%. It continues to grow, but at a more moderate rate, until roughly 45% of the data has been removed. Here, the correlation has an approximate value of .80. From that point it starts to decrease, arguably because of an excessive reduction in the variability of both measures. This analysis indicates that the automatic estimation is not (or is at least less) affected by the familiarity of words. One can thus assume that, for the rare words, DIC–LSA provides an estimate closer to that which raters would give assuming they were familiar with the words.

Generally speaking, even if correlations between the estimated and actual values are high (at least in reference to

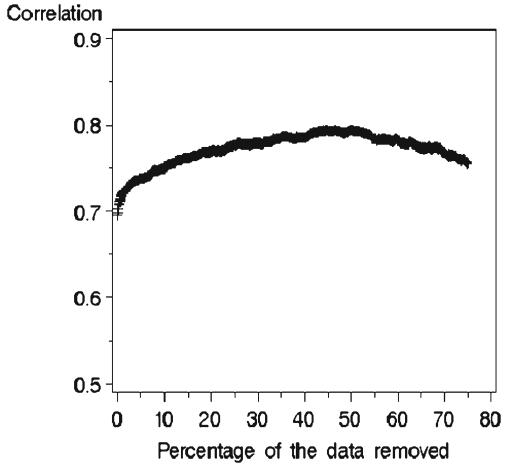

Fig. 1 Correlations between actual imagery scores and DIC-LSA estimates in function of the percentage of the least familiar words removed from the data

concreteness, imagery, and valence), they are far from perfect, and the proportion of explained variance is at most 63%. It should be noted, however, that similar levels of correlations (approximately .70) were considered to be sufficient in the context of using subjective ratings of age of acquisition (AoA) by adults to estimate objective AoA obtained from children's responses to a picture naming task (Bonin, Barry, Méot, & Chalard, 2004; Chalard, Bonin, Méot, Boyer, & Fayol, 2003; Ferrand et al., 2008; Morrison, Chappell, & Ellis, 1997). Moreover, the correlations typically obtained between subjective frequency ratings and objective (corpus-based) frequency measures seldom exceed .70 (Tanaka-Ishii & Terada, 2011; Thompson & Desrochers, 2009). More importantly, in a detailed study of the reliability of imagery ratings for 3,600 French nouns, Desrochers and Thompson (2009) observed a mean test–retest correlation of .73 across sections (min = .28, max = .96) and a mean correlation of .71 between a participant's ratings and the mean ratings of all other participants (min = . 41, max = . 85). These values are just above the correlation between DIC–LSA estimates and the imagery ratings collected by Gilhooly and Logie (1980).

Using DIC–LSA to select words on a dimension

Lexical norms are frequently used in psychology to select experimental materials consisting of words that belong to each of the two extremes of a specific dimension (concrete vs. abstract words; positive vs. negative words). An important question is whether the automatic procedure is effective for this use. To answer this question, the words of each of the five norms were split into two sets based on the rating median. For DIC–LSA (with k = 30), the estimated value was used as a confidence measure that the word would be classified correctly, with decisions regarding words that fell closer to the middle of the spectrum being more uncertain. Several cutoffs were employed: the median split (as for the norms) and the suppression of 25%, 33.3%, 50%, and 80% in the middle of the scale.

Table 2 gives the number of words of the norms categorized by DIC–LSA (N), the percentage accuracy in categorizing the words in one of the two extremes (A), and the value of Cohen's Kappa coefficient (K), a chance-corrected measure of agreement.

As is shown in Table 2, the errors made by the automatic procedure center on the most neutral words. When DIC–LSA is confident about the categorization of certain words, exactitude and kappa are very high. It is noteworthy that the analyses for the imagery norms are based on all words without taking into account the low rates of familiarity of certain words.

The main limitation of the aforementioned analysis is that it was performed on all words present in the specified norms. One can therefore assume that the categorization based on ratings of some of the words is dubious. Upon taking this factor into account, the analyses were repeated by eliminating the 25% most neutral words from the norms. The results, as seen in Table 3, demonstrate a very high efficiency of the automatic procedure.

Qualitative analysis of major discrepancies

The supplementary data files provide the estimated values for each of the five norms for the 17,350 words present in the semantic space. When one of these words is present in the corresponding norms, the human rating value is also provided. Major discrepancies between the rated and the estimated values identify words that merit precise analysis before being included in experimental materials. On the imagery dimension, this index distinguishes words such as *yucca*, *phaeton*, and *underbrush* from words such as *circumstance*, *exception*, and *authority*. The former set of words contains imageable words for the automatic procedure (values ≥ 4.98), but not for the raters (values ≤ 3.13), presumably because of their lack of familiarity for most of the raters. Words in the latter set can be considered as truly not very imageable words. The two procedures afford them scores less than or equal to 3.33.

The analysis of major discrepancies also highlights the impact of polysemy. This phenomenon occurs especially when a word belongs to several grammatical categories, mainly those of noun and verb. In this case, the judges rated the noun—for example, *treat* (7.36 on the valence dimension), *lead* (5.97 on concreteness), and *record* (5.97 on imagery). Conversely, the DIC–LSA estimate was primarily based on the most frequent category in the corpus, often the verb that yielded a value of 3.42 for *treat* on valence, 3.38 for *lead* on concreteness, and 4.37 for *record* on imagery. Other examples occur within a grammatical category. *Scum*,

Table 2 Accuracy of DIC-LSA in dichotomizing dimensions

	Valence			Arousal			Dominance			Concreteness			Imagery		
	N	A	K	N	A	K	N	A	K	N	A	K	N	A	K
0%	950	77	.53	950	71	.41	940	73	.46	1,702	82	.64	1,702	78	.56
25%	712	84	.69	712	76	.52	702	80	.60	1,276	89	.79	1,276	85	.70
33%	632	87	.73	632	79	.58	625	81	.61	1,134	91	.82	1,134	87	.73
50%	474	92	.84	474	83	.67	469	84	.68	850	94	.88	850	90	.79
80%	190	97	.94	190	88	.77	189	90	.80	340	98	.96	340	94	.87

N number of words of the norms categorized by DIC-LSA; *A* percentage accuracy in categorizing the words in one of the two extremes; *K* Cohen's Kappa coefficient. The differences in sample size between norms collected in the same study result from the suppression of words whose score is exactly equal to the threshold

which literally refers to *foam*, a rather neutral meaning with a strong presence in the corpus, scores 5.24 on valence for DIC–LSA while it figuratively means a worthless person, a meaning that corresponds more with the human ratings (2.43). C*rock*, or foolish talk, is less imageable (3.57 in the norms) than its meaning, a pot (5.94 for DIC–LSA).

In other cases, the discrepancies seem to result from the fact that the word is used in contexts that favor an opposite value to that given by the raters as is the case for the word *human*, which is rated as highly concrete (6.14 in the norms), whereas it very often occurs in collocation with abstract words in the corpus, such as *human being* or *human dignity.* This explains why DIC–LSA afforded it a smaller value (3.24). This problem is particularly acute in the case of valence and frequently occurs with words that can be the cause or the consequence of something with the opposite value. For instance, we always *rescue* someone from a negative situation, or we say that someone is *alive* because he or she survived tragic circumstances. The same phenomenon also applies to words such as *debt*, which receives very negative ratings (2.22 in the norms). Simultaneously, this word receives a score of 5.62 in the automatic procedure because it is thematically related to words such as *dollar*, *millionaire,* and *money*, all of which were very positively rated. Another example is *innocent* (6.51 in the norms), which scores 3.05 on valence by DIC–LSA because one finds many words such as *guilty* (cosine = .84, valence = 2.63), *crime* (cosine = .80, valence = 2.89) and *jail* (cosine = .62, valence = 1.95) among its nearest neighbors. This effect can be linked to the fact that the LSA cosine between antonyms is often very high. In the semantic space we built, the cosine between *love* (valence: 8.72 in the norms, 5.71 for DIC–LSA) and *hate* (valence: 2.12 in the norms, 4.09 for DIC-LSA) is .57, hence meaning that *love* is the sixth nearest neighbor of *hate* (after words such as *mad*, *terrible*, and *stupid*) and *hate* the ninth nearest neighbor of *love* (after *affection* and *joy*, but also *jealousy*). Using a sufficient number of neighbors enables DIC–LSA to reduce the estimation error. It is nevertheless crucial that the potential users of the lists that complement the present article be aware of this problem.

In some cases, the discrepancies stem directly from the corpus's specific characteristics. For example, *maggot* (valence: 2.06 in the norms, 4.40 for DIC–LSA) appears often in scientific texts that use objective language to describe the worm ("*A maggot looks like a tiny white worm*"). Similarly, *tomb* (valence: 2.94 in the norms, 5.33 for DIC–LSA) appears almost exclusively in historical texts related to the mortuary practice of the ancient Egyptians, who had a positive vision of death, or in archaeological texts describing rare discoveries.

In some cases, several factors can combine their effect. The polysemy and the type of texts included in the corpus are most likely at the origin of the discrepancy in valence for the word *tragedy*, which refers to a tragic event; therefore, the word assumes a negative connotation (1.76 in the norms) while simultaneously assuming the meaning of a

Table 3 Accuracy of DIC-LSA in dichotomizing norms based on the 75% of the more extreme words in the norms

	Valence			Arousal			Dominance			Concreteness			Imagery		
	N	A	K	N	A	K	N	A	K	N	A	K	N	A	K
0%	706	84	.67	706	75	.49	702	79	.57	1,268	89	.78	1,225	86	.71
25%	561	90	.79	540	81	.63	545	86	.71	1,031	95	.90	972	91	.82
33%	507	92	.83	489	84	.68	492	86	.72	943	96	.92	877	93	.86
50%	404	96	.91	377	88	.76	379	89	.78	741	98	.95	678	95	.91
80%	176	98	.97	165	91	.82	163	96	.91	316	99	.98	300	97	.93

N number of words of the norms categorized by DIC–LSA; *A* percentage accuracy in categorizing the words in one of the two extremes; *K* Cohen's Kappa coefficient

literary genre, a more positive meaning frequent in the corpus (6.20 for DIC–LSA).

Conclusion

We have presented a technique for estimating lexical norms based on the LSA of a corpus. The performed analyses emphasize its effectiveness for estimating concreteness, imagery, and valence. This efficiency was achieved despite the fact that DIC–LSA takes only into account one type of data that humans use to learn semantic representations, that is, the statistical distribution of words in language, and thus neglects a second source of data, that is, the perceived physical properties associated with the referents of words, which, according to Andrews, Vigliocco, and Vinson (2009, p. 463) include affective properties, such as whether something is pleasant or unpleasant. This observation ties in with a series of studies that show that the LSA analysis of word co-occurrences in language can approximate mental representations that have a perceptual origin (e.g., Kintsch, 2007; Louwerse, 2011). The technique is less effective when estimating arousal and dominance; however, these norms are also those for which human judges are the most variable in their ratings.

The procedure has some limitations that are mainly demonstrated in the qualitative analysis of the largest discrepancies between human ratings and automatic estimates. A detailed analysis of these discrepancies would likely make it possible to propose some improvements. For example, it should be possible, based on a tagging preprocessing step, to partially distinguish verbal and nominal forms, thus obtaining independent estimates for *lead* as a verb or as a noun. However, reducing the total number of occurrences of a large number of words would be a side effect and could therefore reduce the efficacy of the procedure. The fact that the texts included in the corpus influence the automatic estimate of certain words suggests that greater efficacy could be achieved by comparing estimates from several corpora; each corpus would act as one of the raters in a normative study. Such an approach would make it possible to study the stability of the estimates and to compute a confidence index in these estimates. Another question that remains unanswered is the minimal size of the norms necessary to obtain an acceptable performance. Norms that contain approximately 1,000 words appear to be sufficient; but can one use less and nevertheless get an efficient estimation?

This technique presents a number of advantages for psychological studies that rest on lexical norms. It can be used to highlight words for which the rating is very different from the estimate. Such a control could be very beneficial for studies that utilize multiple regression analyses that focus on unselected word samples in order to predict the RTs obtained in word recognition experiments. Regarding the extension of norms, a word of caution is essential. Although the analyses reported above indicate that the procedure is generally effective to select extreme words on a dimension, they also show that it can make very serious mistakes for certain words. Moreover, we were able to estimate the effectiveness of the procedure only on the words included in the norms. Since these words do not constitute a random sample of English words, hasty generalizations concerning the effectiveness index for words absent from the norms should be avoided. Exclusive reliance on the values the technique produces to select experimental materials is not advisable. It is by far preferable to use it in order to select candidate stimuli for norming by humans prior to the experiment or for reducing the number of raters, such as is the case in work on automatic essay grading (Warschauer & Ware, 2006). The c. 17,000 words for which DIC–LSA estimates on valence, arousal, dominance, concreteness, and imagery are provided in the supplementary data files are a first step toward this kind of use. They could also be used as a baseline for evaluating future novel approaches that would aim at improving the estimates.

Author note Yves Bestgen is Research Associate of the Belgian Fund for Scientific Research (F.R.S-FNRS). We gratefully acknowledge the help of T.K. Landauer with the TASA corpus.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116,* 463–498. doi:10.1037/a0016261

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133,* 283–316.

Bayer, M., Sommer, W., & Schacht, A. (2010). Reading emotional words within sentences: The impact of arousal and valence on event-related potentials. *International Journal of Psychophysiology, 78,* 299–307. doi:10.1016/j.ijpsycho.2010.09.004

Bestgen, Y. (1994). Can emotional valence in stories be determined from words? *Cognition and Emotion, 7,* 21–36.

Bestgen, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes [Determination of the emotional valence of terms in large corpora]. In: Y. Toussaint, & C. Nedellec (eds) *Actes du Colloque International sur la Fouille de Texte CIFT '02*, 81-94. INRIA, Nancy, France.

Bestgen, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, & D. Tapias (eds) *Proceedings of LREC '08, 6th Language Resources and Evaluation Conference*, 496-500, ELRA, Marrakech, Morocco.

Bonin, P., Barry, C., Méot, A., & Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory and Language, 50,* 456–476. doi:10.1016/j.jml.2004.02.001

Bradley, M. M., & Lang P. J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings (Tech. Rep. No. C-1). Gainesville, FL: Center for Research in Psychophysiology, University of Florida.

Bradshaw, J. L. (1984). A guide to norms, ratings, and lists. *Memory and Cognition, 12,* 202–206.

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. doi:10.1080/17470218.2010.503374

Chalard, M., Bonin, P., Méot, A., Boyer, B., & Fayol, M. (2003). Objective age-of-acquisition (AoA) norms for a set of 230 object names in French: Relationships with other variables used in psycholinguistic experiments, the English data from Morrison et al. (1997) and naming latencies. *European Journal of Cognitive Psychology, 15,* 209–245.

Cohen, A. S., Minor, K. S., Najolia, G. M., & Hong, S. L. (2009). Laboratory-based procedure for measuring emotional expression from natural speech. *Behavior Research Methods, 41,* 204–212. doi:10.3758/BRM.41.1.204

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41,* 391–407.

Desrochers, A., & Saint-Aubin, J. (2008). Sources de matériel en français pour l'élaboration d'épreuves de compétences en lecture et en écriture [Materials for assessing writing and reading skills in French language]. *Canadian Journal of Education*, *31*, 305–326. Retrieved June 22, 2011, from http://www.csse-scee.ca/RCE/Articles/RCE31-2.html

Desrochers, A., & Thompson, G. L. (2009). Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior Research Methods, 41,* 546–557. doi:10.3758/BRM.41.2.546

Dewaele, J.-M., & Pavlenko, A. (2002). Emotion vocabulary in interlanguage. *Language Learning, 52,* 263–322.

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC'06, 5th Language Resources and Evaluation Conference* (pp. 417-422). Retrieved June 21, 2011, from http://sentiwordnet.isti.cnr.it/

Ferrand, L., Bonin, P., Méot, A., Augustinova, M., New, B., Pallier, C., & Brysbaert, M. (2008). Age of acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables. *Behavior Research Methods, 40,* 1049–1054. doi:10.3758/BRM.40.4.1049

Gilhooly, K. J., & Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1,944 words. *Behavior Research Methods and Instrumentation, 12,* 395–427.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114,* 211–244. doi:10.1037/0033-295X.114.2.211

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In: P. Cohen, & W. Wahlster (eds) *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics* (pp. 174-181), Morgan Kaufmann, San Francisco. Retrieved April 5, 2011, from http://acl.ldc.upenn.edu/P/P97/

Heise, D. R. (1965). Semantic differential profiles for 1000 most frequent English words. *Psychological Monographs, 79,* 1–31.

Jessen, F., Heun, R. R., Erb, M. M., Granath, D. O., Klose, U. U., Papassotiropoulos, A. A., & Grodd, W. W. (2000). The concreteness effect: Evidence for dual coding and context availability. *Brain and Language, 74,* 103–112.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114,* 1–32. doi:10.1037/0033-295X.114.1.1

Kamps, J., & Marx, M. (2002). Words with attitude. *Proceedings of the 1st International Conference on Global WordNet,* 332-341. CIIL, Mysore, India.

Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. *Proceedings of COLING '04, the 20th International Conference on Computational Linguistics*, 1367-1373. Geneva, Switzerland. doi:10.3115/1220355.1220555

Kintsch, W. (Ed.) (2007). Meaning in context. In T. K. Landauer, D. McNamara, D. Simon, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 89-105). Mahwah, NJ: Erlbaum.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General, 140,* 14–34.

Kroll, J. F., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 92–107.

Lachenbruch, P. A., & Mickey, R. M. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10,* 1–11.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Landauer, T. K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. *Proceedings of the National Academy of Science, 101,* 5214–5219. doi:10.1073/pnas.0400341101

Landauer, T. K., McNamara, D., Simon, D., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science, 3,* 273–302. doi:10.1111/j.1756-8765.2010.01106.x

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28,* 203–208. doi:10.3758/BF03204766

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3,* 235–244.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology, 50A,* 528–559.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2,* 1–135.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547–577.

Proctor, R. W., & Vu, K. L. (1999). Index of norms and ratings published in the Psychonomic Society journals. *Behavior Research Methods, Instruments, & Computers, 31,* 659–667.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In: D. Jones (ed) *Proceedings of the International Conference on New Methods in Language Processing*, 44-49. University of Manchester, Manchester, England.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, B36,* 111–147.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Tanaka-Ishii, K., & Terada, H. (2011). Word familiarity and frequency. *Studia Linguistica, 65,* 96–116. doi:10.1111/j.1467-9582.2010.01176.x

Thompson, G. L., & Desrochers, A. (2009). Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. *Behavior Research Methods, 41,* 452–471. doi:10.3758/BRM.41.2.452

Tipples, J. (2010). Time flies when we read taboo words. *Psychonomic Bulletin & Review, 17,* 563–568. doi:10.3758/PBR.17.4.563

Turney, P. D., & Littman, M. (2002). *Unsupervised learning of semantic orientation from a hundred-billion-word corpus*. Technical Report ERB-1094 (NRC-44929). National Research Council Canada. Retrieved September 7, 2011, from http://arxiv.org/abs/cs/0212012v1

Turney, P. D., & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21,* 315–346.

Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing affective lexical resources. *PsychNology Journal*, *2*, 61-83. Retrieved September 8, 2011, from http://207.210.83.249/psychnology/index.php?page=psychnology-journal-volume-2-number-1

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10,* 157–180. doi:10.1191/1362168806lr190oa