Behav Res
DOI 10.3758/s13428-015-0639-3

Relative meaning frequencies for 578 homonyms in two Spanish dialects: A cross-linguistic extension of the English eDom norms

Blair C. Armstrong[1] · **Camila Zugarramurdi**[2] · **Álvaro Cabana**[2] · **Juan Valle Lisboa**[2,3] · **David C. Plaut**[4]

Abstract Relative meaning frequency is a critical factor to consider in studies of semantic ambiguity. In this work, we examined how this measure may change across the European and Rioplatense dialects of Spanish, as well as how the overall distributional properties differ between Spanish and English, using a computer-assisted norming approach based on dictionary definitions (Armstrong, Tokowicz, & Plaut, 2012). The results showed that the two dialects differ considerably in terms of the relative meaning frequencies of their constituent homonyms, and that the overall distributions of relative frequencies vary considerably across languages, as well. These results highlight the need for localized norms to design powerful studies of semantic ambiguity and suggest that dialectal differences may be responsible for some discrepant effects related to homonymy. In quantifying the reliability of the norms, we also established that as few as seven ratings are needed to converge on a highly stable set of ratings. This approach is therefore a very practical means of acquiring essential data in studies of semantic ambiguity, relative to past approaches, such as those based on the classification of free associates. The norms also present new possibilities for studying semantic ambiguity effects within and between populations who speak one or more languages. The norms and associated software are available for download at http://edom.cnbc.cmu.edu/ or http://www.bcbl.eu/databases/edom/.

Keywords Semantic ambiguity · Homonyms · Cross-linguistic/dialect differences · Rating dictionary definitions · Norm reliability

Given that the vast majority of words are semantically ambiguous—that is, their meanings depend on the context in which they occur—a comprehensive theory of word and discourse comprehension necessarily involves a theory of semantic ambiguity resolution (Klein & Murphy, 2001, 2002). In studies of semantic ambiguity, homonyms represent one theoretically important type of item: Single word forms that are associated with two or more unrelated interpretations (e.g., the <river> and <money> interpretations of BANK, hereafter referred to in the form <river>/<money> BANK). Across a series of investigations, as compared to relatively unambiguous words such as CHALK, homonyms have been reported as showing an overall processing advantage (Hino, Pexman, & Lupker, 2006; although see Armstrong & Plaut, 2011, for discussion), neither a disadvantage nor an advantage (e.g., Rodd, Gaskell, & Marslen-Wilson, 2002) or a processing disadvantage (e.g., Mirman, Strauss, Dixon, & Magnuson, 2010). Although the theoretical debate regarding the source of all of these discrepancies is ongoing (see, e.g., Armstrong & Plaut, 2008, 2011; Hino, Kusunose, & Lupker, 2010; Hino et al., 2006; Rodd et al., 2002), there is general agreement on one point in this literature: the relative frequency of a homonym's interpretations can modulate the effects of homonymy (e.g., Armstrong & Plaut, 2011; Klepousniotou, Pike, Steinhauer, & Gracco, 2012; Klepousniotou, Titone, & Romero, 2008; Mirman et al., 2010; Seidenberg, Tanenhaus, Leiman, & Bienkowski,

✉ Blair C. Armstrong
blair.c.armstrong@gmail.com; b.armstrong@bcbl.eu

[1] Basque Center on Cognition, Brain, and Language, San Sebastian-Donostia, Spain

[2] Facultad de Psicología, Universidad de la República, Montevideo, Uruguay

[3] Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

[4] Psychology Department and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA

Published online: 15 August 2015

1982; Swinney, 1979; Tabossi, 1988). Consequently, quantifying the relative meaning frequency of a homonym plays a critical role in contextualizing any effects obtained with this type of item and in determining the broader implications for theories of semantic ambiguity resolution.

To this end, Armstrong, Tokowicz, and Plaut (2012) introduced the eDom norming software and relative meaning frequency norms for American English and provided evidence that their methods produce more reliable and valid estimates of relative frequency than prior, more resource-intensive techniques such as estimations based on free association norms (e.g., Twilley, Dixon, Taylor, & Clark, 1994). In brief, as is exemplified by the screenshot presented in Fig. 1, the eDom software package enables the rapid collection of relative frequency norms by presenting participants with the dictionary definition for each meaning of a homonym and allowing them to indicate the relative frequency with which the word is used to denote each meaning. Participants also have the option to input their own definitions for other meanings of the word that are not covered by the dictionary definitions (this can be done in the yellow boxes illustrated in the figure) and rate the relative frequency with which those meanings are encountered. The resulting norms have several advantages over past methods, such as the classification of free associates: Individual participants can complete a norming session very rapidly, no time-consuming and subjective classification of their responses must be completed by a group of raters, an order-of-magnitude fewer participants must rate each item before stable and reliable norms can be computed (a point that is examined in additional detail in the present work), and the resulting norms are better predictors of behavioral performance in tasks such as lexical decision.

The success of the original eDom study raises the possibility that analogous studies of semantic ambiguity in other languages, such as Spanish, would likely benefit from their own sets of eDom norms. However, one question that the initial work did not answer was whether one set of norms would be sufficient for use across the various dialects of a language, or whether there is nontrivial variability in relative meaning frequencies across dialects. To date, no study has, to our knowledge, addressed this issue in detail. However, two observations identify this as a worthy target for investigation, over and above the surface validity of presuming that sociocultural differences could shape relative meaning frequency norms. The first is that cross-dialectal variation in the relative frequencies of some homonyms (e.g., <crack>/<man> CHAP) may have been responsible for some discrepant results obtained with the same homonyms in British versus American English (Armstrong & Plaut, 2008, 2011; Beretta, Fiorentino, & Poeppel, 2005; Rodd et al., 2002). The second is that there are significant differences between British- and American-English-localized word frequency norms.[1] Although detailed methodological differences and the nature of the base corpora used to derive the frequency norms might account for some of this difference, dialect effects do show some possible external validity in terms of small but significant improvements in how well dialect-localized word frequency data predict the lexical decision latency and accuracy in the British English versus the American English Lexicon Project.[2] If word frequencies and their resulting effects on performance can noticeably change between dialects, it stands to reason that relative meaning frequencies could change, as well.

Given these observations, in the present work we aimed to expand upon previous work in several important ways, using a Spanish version of the eDom software to study the relative meaning frequencies associated with homonymous words in the Rioplatense[3] versus the European dialects of Spanish. This work shows that the eDom norming methods are sensitive to relative meaning frequency differences across the two dialects. In so doing, the practical value of collecting regionally localized and up-to-date relative frequency estimates is established, providing quantitative evidence that some weak and inconsistent effects observed using nearly identical sets of materials and methods may have been due, at least in part, to dialectical differences in English (see, e.g., Armstrong & Plaut, 2008, 2011; Beretta et al., 2005; Rodd et al., 2002). Additionally, detailed analyses of the interrater reliability

[1] The correlation between the log-transformed British National Corpus (BNC) word frequency data, which were derived from a cross-sampling of written and spoken input, and the equivalent data from the SUBTL word frequency data for American English, which was derived from film and television subtitles, was .81. All 18,545 words for which correct latency and accuracy information was available in the (restricted) American English Lexicon Project (ELP) and the British English Lexicon Project (BLP) were included in this calculation (Balota et al., 2007; Brysbaert & New, 2009; Keuleers, Lacey, Rastle, & Brysbaert, 2012).

[2] Log-transformed BNC and SUBTL word frequency data were used to predict accuracy and correct latencies for all 18,545 trials common to both the BLP and the ELP lexical decision data. The results showed, with only one numerically consistent but statistically marginal caveat, that correlations were significantly stronger when the dialects of English associated with the word frequency norms and the lexical decision data matched (Correct latencies: SUBTL–ELP $r = -.630$ vs. BNC–ELP $r = -.595$, $p < .01$, one-tailed; SUBTL–BLP $r = -.641$ vs. BNC–BLP $r = -.650$, $p = .07$; Accuracy: SUBTL–ELP $r = .482$, BNC–ELP $r = .461$, $p < .01$; SUBTL–BLP $r = .509$, BNC–BLP $r = .542$, $p < .01$). Similar results were also obtained when only items with word frequencies less than 100 were included, to avoid the nonlinear effects of log-transformed word frequency above that level (Brysbaert & New, 2009). These results do, however, contrast with the SUBTL-only frequency comparisons of a much smaller subset of items between the two corpora, which showed stronger correlations between SUBTL and the BLP (Keuleers et al., 2012).

[3] Rioplatense Spanish is the dialect of Spanish spoken primarily in areas surrounding the Rio Plate, primarily Uruguay and Argentina (predominantly in Buenos Aires, Patagonia, and the Argentine Littoral).

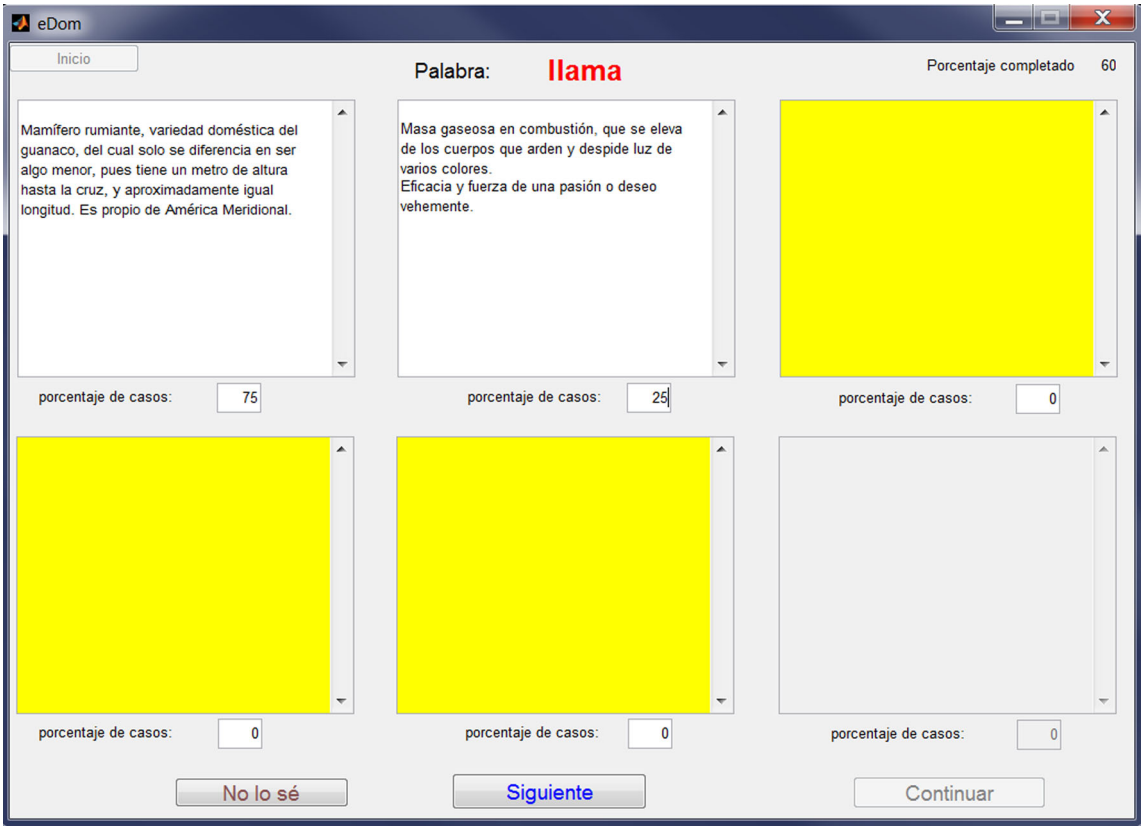

Fig. 1 Screen capture of the Spanish version of the eDom norming software during the norming of the word LLAMA

and the rate at which stable relative meaning frequency estimates were obtained show that the eDom norming method is even more efficient than previously claimed in terms of the total number of observations needed to generate a set of norms, making it very feasible to develop region-localized norms as part of standard semantic-ambiguity research projects. More generally, the availability of large sets of relative frequency norms and methods in Spanish as well as in English—two of the most frequently spoken Indo-European languages—opens up the possibility for additional cross-linguistic comparisons and investigation of related phenomena such as translation ambiguity (e.g.,. Degani, & Tokowicz, 2013; for a review, see Degani & Tokowicz, 2010), while simultaneously addressing criticisms regarding the Anglocentrism and uncertain universality of much recent psycholinguistic research (Carreiras, Armstrong, Perea, & Frost, 2014; Frost, 2012; Lerner, Armstrong, & Frost, 2014; Share, 2008).

Experiment

In the experiment, we assessed the efficiency, reliability, and cross-linguistic similarity of relative meaning frequency norms derived from dictionary definitions and supplemented by participant-generated definitions in Rioplatense and European Spanish.

Method

Participants

Europe A total of 95 participants completed the experiment (63 female, 32 male; mean age = 22.3, *SE* = 0.3). The participants were exposed to Spanish from a very early age (*M* = 0.76 years, *SE* = 0.2, max = 7) and showed very high overall proficiency in Spanish, as assessed using the BEST test (an adaptation of the MINT multilingual naming task for Basque, English, and Spanish; Duñabeitia, Casaponsa, Dimitropoulou, Martí, Larraza, & Carreiras, 2014; Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2012; *M* = 76.1/77, *SE* = 0.2) and via in-person interviews that assessed fluency (*M* = 4.96/5, *SE* = 0.02). Participants were recruited at the University of the Basque Country in San Sebastián (UPV/EHU–Donostia), and were either completing or had recently completed an undergraduate degree. All of the participants were bilingual, to varying degrees, in Basque, and many had also been exposed to a third language (e.g., English, French, German). Knowledge of other languages did not appear to directly influence the relative meaning frequency data,

however, as determined by the lack of new definitions being generated by the participants that were associated with other languages. Participants were paid for their contributions to the investigation.

Rioplatense A total of 149 students enrolled in Psychology courses at Universidad de la República in Uruguay took part in the study. All participants self-reported as native speakers of Spanish. Consistent with Uruguayan law on research with humans, participation was entirely voluntary and no remuneration was provided. For this reason, some simple modifications of the task were made to reduce its overall length and to encourage participation, such as reducing the total number of trials and collecting less extensive demographic data. Basic demographic data were available for a random subsample of 27 participants (mean age = 25.4, *SE* = 4.1; 14 female, 13 male).

Stimuli

Similar to the prior study in English, the main experimental stimuli consisted of a large sample of homonyms selected so as to satisfy standard constraints on experimental items in psycholinguistic research, and to obtain a set that was comparable in size to the English eDom norms. These items were drawn from an exhaustive set of 1,857 homonyms and homophones identified via an automated parsing of the dictionary of the Real Academia Española (RAE) (2001), the official dictionary for European Spanish, which has been extended to include definitions from South American dialects, as well. Using supplementary psycholinguistic data obtained from the EsPal database (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013), this list was filtered down to contain entries that had a written word frequency between 1 and 100 (and one word with a frequency of 160); a length, in letters, between 3 and 10; two or more unrelated definitions in the RAE dictionary, and at least one sense corresponding to a noun, adjective or verb definition. Counts of the number of related senses for each of the unrelated meanings of the homonym were obtained by summing the number of definitions listed within the entries for each unrelated meaning. Grammatical class counts (e.g., number of nouns vs. number of verb interpretations) associated with each homonym were calculated by summing the grammatical classes associated with the different interpretations of a word across all of its meanings. These methods of identifying a relatively exhaustive set of unrelated meanings and of measuring the total number of related senses and grammatical classes associated with a word have already been established in English (Armstrong, Tokowicz, & Plaut, 2012; see also Azuma & Van Orden, 1997; Rodd et al., 2002). This screening identified 663 homonyms for use in the norming study. The majority of the selected items had either two meanings (522 items) or three meanings (119). As an extension of the original work, 133 of the homonyms that were included in the set were also homophones (e.g., the homonym <hunt>/<fabric> CAZA is also a homophone of <house> CASA in Spanish), so as to provide some normative data that would be useful for assessing the relationship between relative meaning frequency and homophony (e.g., as an extension of Seidenberg & McClelland, 1989). No homographs were included because this class of items effectively does not exist in orthographically transparent languages such as Spanish. An additional ten items were included in the Rioplatense data to collect norms for items used in a prior experiment, and were excluded from all subsequent analyses.

Procedure

Before beginning the experiment, participants were given a briefing covering what they needed to do in the task that was a direct translation of the instructions used in the English eDom experiment. Each participant then rated a randomly selected subset of the total set of experimental items.[4] Participants from Europe rated approximately 110 items, whereas the Rioplatense participants rated approximately 42 items. The possible impact of this difference is assessed in the results section. Factoring in the total number of participants in each dialect, this led to the collection of approximately 16 ratings for each item in the European dialect and nine ratings per item in the Rioplatense dialect. A Spanish version of the eDom software was used to collect the relative meaning frequencies, and is available on the eDom website (http://edom.cnbc.cmu.edu/edomnorms.html). This software presents all of the dictionary definitions of each unrelated meaning of a homonym on the screen, one at a time, in a random order, and provides space for participants to list additional definitions that they know that do not appear in the dictionary definitions. Participants then indicate, in percent, how frequently each of those meanings is denoted when they encounter a given homonym. Participants were able to take self-paced breaks after blocks of approximately 20 words. European participants completed the experiment using desktop computers in standard soundproof behavioral cabins at the Basque Center on Cognition, Brain, and Language. Rioplatense participants completed the experiment in a quiet computer laboratory containing multiple desktop computers and on laptops set up "in the field" in quiet areas of the university that were frequented by undergraduate students.

[4] In the European dialect, the items were counterbalanced across participants, such that each item was seen equally often. A mix of counterbalanced and completely random sampling was employed in the Rioplatense dialect, such that a minimum number of eight data points were available for each item, but some items were sampled more often.

European and Rioplatense participants completed the experiment in approximately 35 min and 15 min, respectively.

Results and discussion

Initial screening

The data from six Rioplatense participants were dropped because they did not complete the full set of ratings assigned to them. Participants were then screened separately in each dialect to eliminate individuals that did not know an atypically large number of words, as determined using the one-tailed *z*-score value associated with $p < .05$. This eliminated two participants from the European group and eight participants from the Rioplatense group, who on average indicated they did not know more than one third of the presented items. For the remaining participants, 11 % of the total responses in the European group and 13 % of the total responses in the Rioplatense group indicated that an item was unknown. The percentage of items that participants indicated they did not know increased fractionally throughout the experiment (on average, 2.0 % of the total "unknown" responses were made in the first quartile vs. 3.3 % in the last quartile).

Dialectal differences in known word forms

Virtually all of the responses indicating that a word was not known were associated with a particular subset of the words, as determined by examining the number and quantity of unknown responses associated with words that were consistently rated as "unknown" by at least 20 % of raters. This analysis showed that 115 items in European and 141 in Rioplatense Spanish were responsible for 9 % and 11 % of the total "unknown" responses in each population, with 84 items being rated as "unknown" above the 20 % threshold in both dialects (see Fig. 2). The fact that the number of words that were unknown in only one of the two dialects (88 in total across both dialects) was similar to the number of words that were unknown in both dialects (84) provides the first piece of evidence that there was nontrivial dialectal variation in participants' knowledge of the rated words. This was also reflected in the degree of correlation between the percentages of participants who knew the words in each language ($r = .53$). Both of these observations are in general agreement with the observed differences between British and American English noted in the introduction. Inspection of the items that were well known in one dialect but not in the other revealed that these differences had plausible sociocultural bases, as have been flagged in recent versions of the dictionary. For instance, the word GIL was well known by all Rioplatense participants but by less than a third of European participants (for whom GIL is most commonly encountered as a family name). This made sense after inspecting the relative meaning frequency data from the Rioplatense dialect, which showed that 97 % of the relative meaning frequency for that word was loaded onto a meaning related to tango music. On the basis of these results, all of the words that were unknown by at least 20 % of participants in both dialects, as well as one item that did not have any "known" ratings in one dialect, were dropped from further analyses. This left a total of 578 homonyms in the set.

Given that the proportion of "unknown" responses in Spanish (12 %) was four times larger than in the original eDom study in English (3 %), we also evaluated whether the distributions of word frequency data, a key predictor of an individual's overall familiarity with a word, may have differed across the two languages. In the English version of eDom, the final set of items after filtering had a mean word frequency of 15.7 per million ($SE = 0.9$), as assessed using word frequency data from television and film subtitles (Brysbaert & New, 2009). In Spanish, the frequency data[5] for the items that were eliminated were considerably lower (mean = 5.4, $SE = 0.7$), but were based on counts from a corpus of written materials. To ensure that the nature of the frequency data was not a confounding factor, and because of the better predictive validity from subtitle counts (Brysbaert & New, 2009), we opted to use the subtitle word frequency data in all of the subsequent analyses, which were available for all but two of the items. Reinspecting the normed items with this alternate measure of word frequency, we found that although the average frequency of the unknown items was similar ($M = 4.9$, $SE = 2.4$), the variability was considerably higher, and 65 % of the unknown words had word frequencies below 1 ($M = 0.36$ $SE = 0.01$). This was likely a strong contributing factor to the higher proportion of "unknown" responses.

Dialectal differences in the meanings captured by dictionary definitions

On average, the sum of the relative frequency ratings for the two most frequent meanings of the items was 96 % in both European Spanish and Rioplatense Spanish, indicating that most meanings are captured by the dictionary and most homonyms effectively only have two meanings for the participants. Despite this strong coverage, however, a new definition for a word was listed on 10 % of trials in the European group and 7 % of trials in Rioplatense group. A nonidiosyncratic definition missing from the RAE dictionary was identified

[5] To our knowledge, the EsPal word frequency data (Duchon et al., 2013) represent the largest and most up-to-date set of word frequencies for Spanish, but they do not distinguish between Rioplatense and European Spanish dialects. Given that no comparably large-scale dialect-specific word frequency norms were available for the Rioplatense dialect and that the main aim of these analyses did not bear on dialect-specific issues, these data were used for all analyses in both dialects.

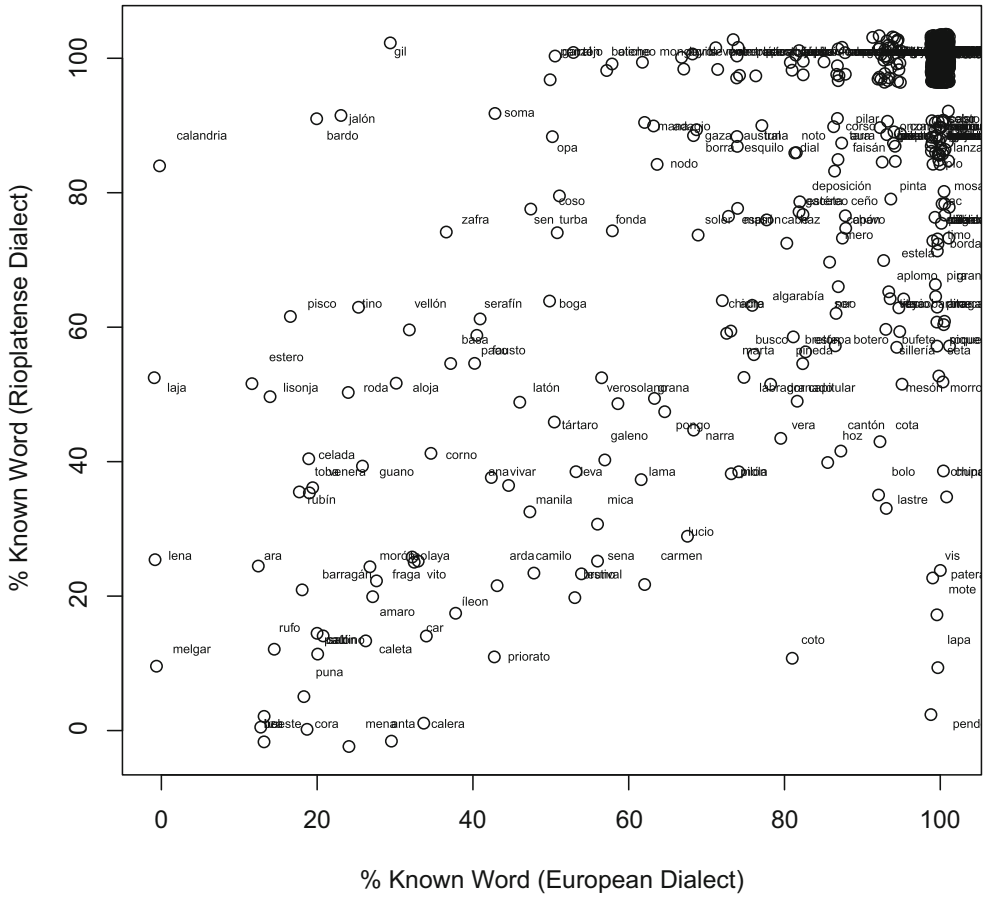

Fig. 2 Plot of the percentages of raters who indicated that they knew the rated word in each dialect. The data have been jittered to highlight that most words were well known in both dialects

whenever a common definition was listed by 40 % of participants in a given dialect. After excluding new definitions that were closely related to part-of-speech extensions of the base meaning (e.g., a new definition for the noun meaning of <error> TACHA denoted the action of committing an error) and two common Spanish names, this led to the identification of six new definitions in European Spanish and 16 new definitions in Rioplatense Spanish. Three of these definitions were common to both dialects. The mean of the largest meaning frequency for these new definitions was 71 %. This indicates that the new definition that was added by participants is generally the dominant meaning of the word. These results suggest that the same approach used to norm English homonyms—starting from an initial set of dictionary definitions and supplementing these definitions with participant-generated definitions—provide very good coverage of the different meanings that are associated with a given word. The results also highlight that the RAE dictionary, although it has recently focused on improving coverage of Latin American interpretations of words, is still missing relatively more word meanings from Latin American dialects, notwithstanding that the dictionary does capture the vast majority of a word's meanings in both dialects. Finally, the high degree of convergence on a small number of definitions suggests that this method is a sensitive means of identifying relatively frequent meanings that are not included in the dictionary.

Comparison of the relative meaning frequencies across dialects

Figure 3 plots the relative meaning frequency of the first dictionary definition for each of the homonyms in the two dialects (similar results were also obtained by simply plotting the largest relative meaning frequencies instead). The shared variance across the two dialects was high (R^2 = 72 %) but still showed considerable dialectal variation. Inspection of the items confirmed that many of these distinctions had plausible cultural and/or linguistic bases. For instance, an example of a culturally driven dialect difference is the word CUCO, which has one meaning that relates to a mythical being in Rioplatense. This meaning receives a high rating in the Rioplatense dialect (99 %) and a considerably lower rating in the European dialect (8 %). Similarly, the word CHUCHO denotes both <dog> and

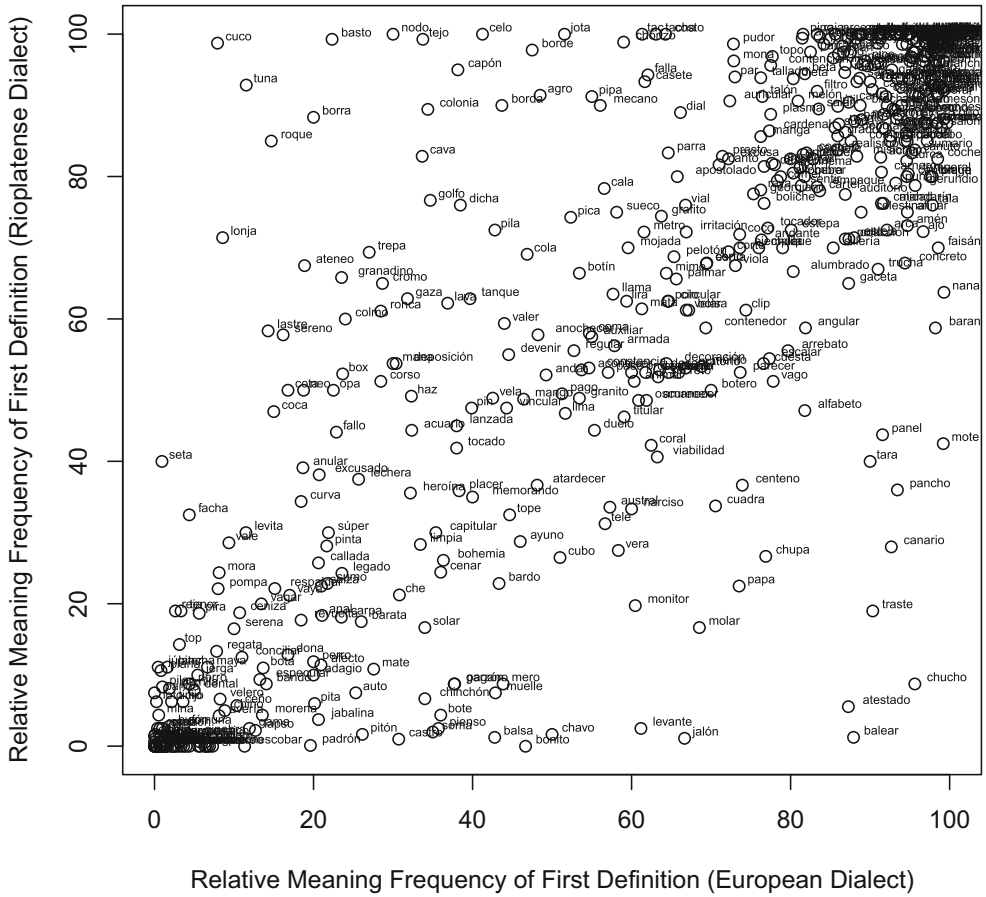

Fig. 3 Relative meaning frequencies for the first dictionary definition for each item in each dialect

<cold> meanings. The <dog> meaning is commonly used in European Spanish (96 %) but rarely used in the Rioplatense dialect (1 %). Collectively, these results also support the notion that dialect differences are important and can be detected by eDom, and could similarly be responsible for some of the inconsistent results obtained in American versus British English using the same items (e.g., Armstrong & Plaut, 2008, 2011; Beretta et al., 2005; Rodd et al., 2002).

Distribution of largest relative-frequency ratings

The degree to which homonyms have relatively balanced (i.e., near 50, for words with two meanings) versus unbalanced (i.e., near 100) relative meaning frequency ratings provides insights into what proportion of words are effectively homonymous in a given language, and to what degree those homonyms might be expected to generate the strong competitive dynamics between relatively balanced interpretation frequencies that are expected by some theories (e.g., Armstrong & Plaut, 2011; Klepousniotou et al., 2008; Mirman et al., 2010; Piercey & Joordens, 2000). To a first approximation, the English literature suggests that homonyms with their largest relative meaning frequencies below 65 % can be treated as balanced homonyms. There is no equivalent accepted standard for when a homonym's interpretations are so unbalanced that one meaning is basically unknown, and therefore the homonym should be treated as being effectively unambiguous. However, homonyms with their largest relative meaning frequencies in excess of 95 % are highly likely to fall into this category, and Armstrong and Plaut (2011) found that even homonyms with relative meaning frequencies above 75 % showed substantially reduced competitive dynamics.

Figure 4 plots the distributions of the largest relative meaning frequencies for the two Spanish dialects and from the original eDom norms collected in English. Overall, these results indicate that both Spanish dialects contain considerably fewer balanced homonyms and considerably more unbalanced homonyms (i.e., effectively unambiguous items); only 7 % of items in the European dialect, and 5 % of the items in the Rioplatense dialect, would be considered to be balanced homonyms. These results have two possible implications for cross-linguistic comparisons of ambiguity effects. One possibility is that fewer relatively balanced homonyms exist in Spanish. If true, this would indicate that ambiguity studies conducted in Spanish that do not control for relative meaning frequency would be

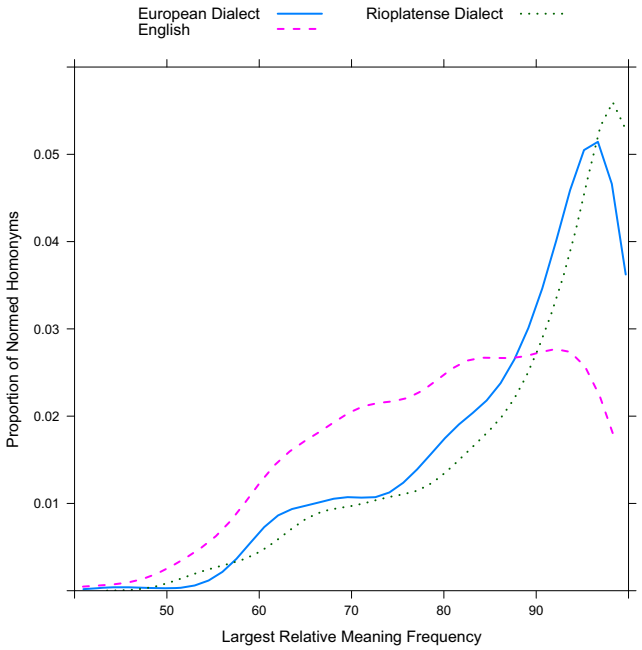

Fig. 4 Proportions of normed homonyms as a function of the largest relative meaning frequencies for English, European Spanish, and Rioplatense Spanish

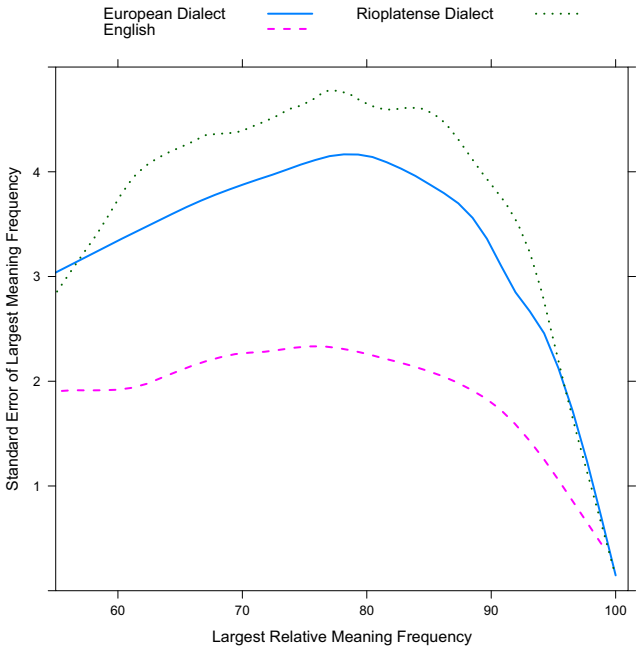

Fig. 5 Standard errors of the largest relative meaning frequencies as a function of the largest relative meaning frequencies for English, European Spanish, and Rioplatense Spanish. A small number of items with their largest relative meaning frequencies below 55 % are not shown

less likely to show different patterns of effects for homonyms relative to unambiguous controls than in a language such as English. Alternatively, despite using a translation of the English instructions to run the eDom norming study in Spanish, it is possible that social, cultural, or other systematic differences between the populations of participants may have caused a systematic bias in the ratings. For instance, the dominant meanings of an English homonym and a Spanish homonym that are equally unbalanced in terms of how often each of their interpretations are actually encountered may receive different relative meaning frequency ratings because one population is more willing to produce more extreme ratings. Strong inferences in this regard will require additional experimental work using unbalanced and balanced homonyms that are carefully matched across languages (see Armstrong, Watson, & Plaut, 2012, for methods relevant to this end).

Reliability

Variability in relative meaning frequency estimates Overall, the variability in the largest relative meaning frequency ratings, as assessed via the standard errors of the means, was quite low in the two Spanish dialects, and only slightly larger than that obtained in English (English mean *SE* = 1.9, European Spanish = 2.7, Rioplatense Spanish = 2.6). This indicates that the normed data should be highly stable across languages and dialects, and that relatively little extra variance is added by having fewer total observations and having more participants rate fewer items, as was the case in Rioplatense versus European Spanish. Figure 5 provides additional insight by plotting variability in relative meaning frequency estimates as a function of relative frequency. Similar to English, both European and Rioplatense Spanish showed the most variability in estimates related to unbalanced homonyms, in the range of 70 %–90 %. This further supports the notion that a simple measure of relative meaning frequency, such as the largest relative meaning frequency of a word, is to be preferred over more complex measures of uncertainty based on information theory, which are associated with greater sensitivity in the 70 %–90 % range and less sensitivity above 90 % (cf. Twilley et al., 1994; see Armstrong, Tokowicz, & Plaut, 2012, for a discussion). Variability in this range was also slightly higher than that observed in English, in part because there were fewer observations in this subsection of the distribution.

Reliability of norms across participants Another understudied issue in the literature is the degree to which individual participants produce similar ratings for a given homonym, and thus, the degree to which individual differences in relative meaning frequency values could have shaped the results of studies focused on mean performance across participants (for additional discussion of the importance of considering individual differences and not only group averages, see Frost, Armstrong, Seigelman, & Christiansen, 2015). To gain insight into this issue, the ratings from each participant were correlated with the average rating across participants. This procedure is analogous to other related methods for assessing interrater reliability by conducting a factor

analysis and examining the degree to which each participant loads on the first factor (Stemler, 2004). However, it does not suffer from the fact that there is, on average, low overlap in the number of items rated by individual pairs of participants if participants rate only a small subset of the total item set, thus leading to a sparsely populated item-by-participant matrix that is unsuitable for factor analysis. Similarly, this approach addresses issues with some classic methods for assessing interrater reliability when agreement levels are high, and the associated adjusted reliability measures thus are more complex (Gwet, 2008). The results indicated a reasonable degree of similarity in the ratings obtained across participants (European Spanish mean r = .69, SE = .01, range = .42–.82; Rioplatense mean r = .44, SE = .01, range = .11–.64). The degree of similarity was slightly lower in the Rioplatense data, possibly because of the additional variability introduced by factors such as having participants rate fewer items and using the mix of counterbalanced and random sampling, as we noted in the Method section.

To determine whether the less-than-perfect similarity between individual participants and the mean ratings was due to qualitative differences between subpopulations of the raters in each dialect, in an additional analysis we recomputed the mean ratings after having dropped the 10 % of participants with the lowest correlations with the mean ratings in the first analysis. The correlation between the initial set of mean ratings and the set of mean ratings that excluded those participants was still extremely high (r > .99 in both dialects). This suggests that the variability in how well individual participants' ratings resemble the mean ratings is primarily due to random noise and not to a systematic deviation amongst subgroups of raters—an issue that could be assessed in future work by retesting the same participants at a later date.

Norm reliability as a function of sample size The results from the previous section indicate that dropping 10 % of the total participants—those with the lowest correlation with the average rating—did not meaningfully change the relative frequency norms, as assessed via the correlation between the pre- and postdropped item means. In light of this outcome, it is worth asking just how many observations, in fact, are needed to achieve reliable norms. One valuable contribution from the first eDom study was that it showed, via both internal measures of reliability and assessments of external validity, that stable and useful norms had been achieved with approximately 16 ratings per item, as opposed to the approximately 100 ratings per item needed using methods based on the classification of free associates. However, that investigation did not establish in detail whether 16 observations was just barely sufficient or was clearly more than necessary to achieve those ends. This issue was investigated in more detail in what follows.

In the first analysis, we assessed how quickly the largest relative frequency ratings stabilized by correlating the mean item ratings obtained with n participants with those obtained with n + 1 participants. Only the items rated by the new participant added to the set were correlated across the two sets, given that those are the only ratings that could change. For each sample size, this calculation was repeated 1,000 times. The sample sizes had a lower bound of ten to avoid situations in which very few items rated by the new participant had previously been rated. These calculations were completed for three different data sets: the European data set, the Rioplatense data set, and a European data set that was restricted to only contain the data from the first 42 items rated by the participants (labeled the "first 42" set in the plots). This allowed for direct comparisons between the reliability of the Rioplatense data and the restricted European data that were not influenced by the increased number of items rated by the European participants. Because the participants were sampled at random, some items could be rated by more participants than other items for a given sample, whereas complete counterbalancing in the norming study ensured that each item was rated equally often for a given sample size. Thus, the results are best interpreted as a lower bound on the reliability function. Via the central limit theorem, it was also to be expected that, eventually, adding more participants—even if their ratings were not correlated with one another—would not meaningfully alter the item ratings. To quantify the degree to which the item means were stabilizing due to the similarity of participants' responses, over and above the stabilization that occurred due to the central limit theorem, a set of "control" functions was also included as part of the analyses, in which each participant's ratings were replaced with random ratings sampled from a uniform distribution across the range [0, 1].

The results are plotted in Fig. 6 and show that the norms stabilized surprisingly quickly. For the European data, the correlation between a set of item means from n participants and from n + 1 participants was already above .9 with only ten participants (approximately two ratings per item). This correlation had effectively reached asymptote after 50 participants' worth of data (five ratings per item) and only increased fractionally by averaging in an additional 25 participants' worth of ratings. The Rioplatense data showed less strong agreement—the analogous correlation for these data did not exceed .9 until 40 participants were tested, and it did not approach an asymptotic value until 125 participants were tested. However, this reduced level of agreement for a given sample size in the Rioplatense data was primarily due to the reduced number of ratings being entered into the analysis. This is illustrated in the plot by the similar (although slightly higher) correlations in the analysis that only included the first 42 trials from the European data. Needless to say, in all of these analyses, the actual data showed substantially more agreement than the control data.

Given the results from the first analysis, it appeared likely that the total number of observations per item, regardless of

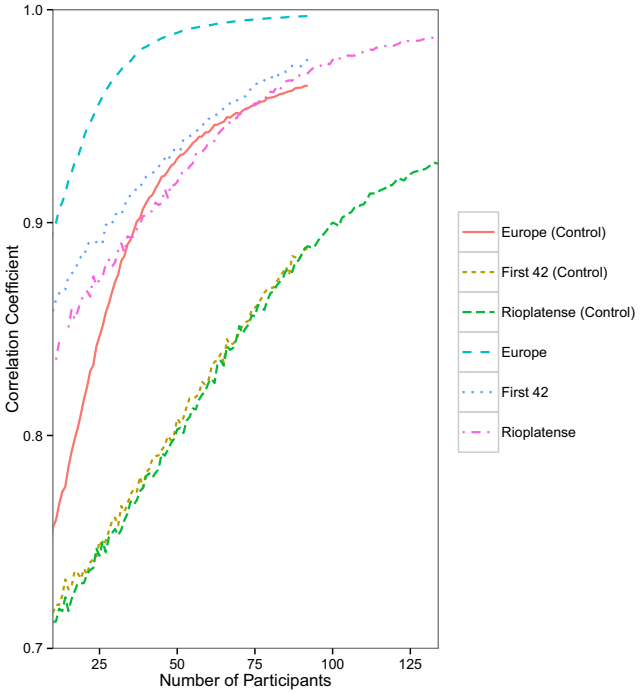

Fig. 6 Correlations between the mean item ratings for *n* participants and the mean item ratings for *n* + 1 participants, as a function of the number of participants. Because of differences in the numbers of items rated by each participant (European vs. Rioplatense) or in the maximum number of participants that had rated an identical number of items (Rioplatense vs. the first 42 ratings from the European participants), the termination point of each function is different. Control functions are included for reference, as described in the main text. Standard errors for these functions are not plotted because they were very small (largest *SE* = .015)

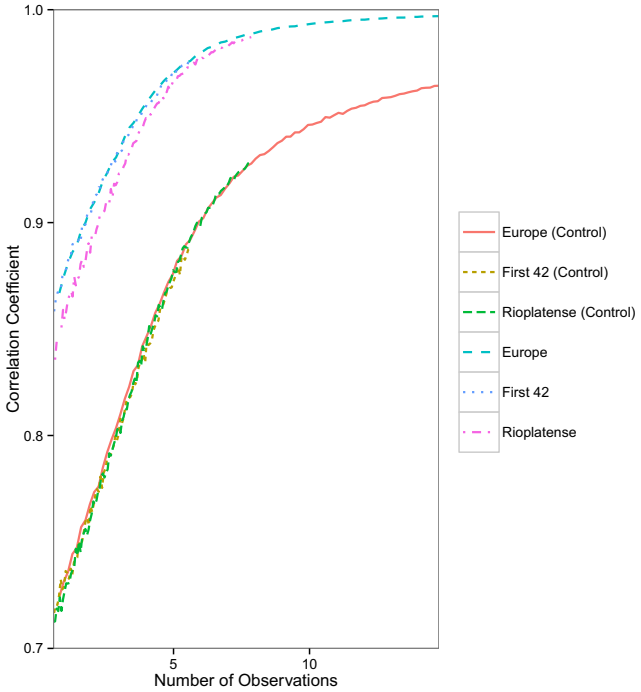

Fig. 7 Correlations between the mean item ratings for *n* participants and the mean item ratings for *n* + 1 participants, normalized to the total number of observations per item. Control functions are included for reference, as described in the main text

whether few participants rated many items or many participants rated few items, was what determined how quickly the norms stabilized. To evaluate this more directly, Fig. 7 plots the same correlation coefficients from the previous plot as a function of the total number of observations per item (i.e., it restandardizes each set of coefficients to equate the numbers of ratings instead of the numbers of participants for each data set on the *x*-axis). This plot shows that, regardless of whether participants rated 42 or 110 items, the correlation coefficient was already above .9 with only three ratings per item, and had effectively reached an asymptotic value near 1.0 with only seven observations per item. These results clearly highlight the modest investment that researchers must make when using computer-assisted norming methods to obtain sensitive region-localized relative meaning frequency estimates. They also illustrate the flexibility that is available in terms of how many items a given participant rates and the conditions under which testing can occur.

Correlation between the largest relative meaning frequency and other psycholinguistic variables We compared the relationship between the largest relative meaning frequency and several semantic, grammatical, lexical, and sublexical predictors to determine whether relationships were observed in Spanish similar to the ones observed in English. Correlations were significant ($p < .05$; two-tailed) in these analyses if the correlation coefficient was greater than .10, given the total number of observations in each statistical test. Marginal effects ($p < .10$) are also indicated below. In each analysis, the equivalent correlation from the English eDom study is indicated for reference (denoted as r_E). These analyses were conducted on the data from the EsPal database, which is primarily aimed at providing psycholinguistic measures for European Spanish, because detailed psycholinguistic data for many of the other measures are not currently available for the two dialects separately. Thus, the following results are likely more representative of the relationships that exist between the largest relative meaning frequency and European Spanish.

At the semantic level, significant correlations were observed between the largest relative meaning frequency data and the number of unrelated meanings ($r = –.22$; $r_E = –.32$) and number of related senses ($r = –.19$; $r_E = –.28$) associated with each word. At the grammatical level, the relationship between the largest relative meaning frequency and the number of verb, noun, and adjective interpretations, collapsed across meanings, was significant only for verbs ($r = –.11$; $r_E = –.26$), not for nouns ($r = –.06$; $r_E = –.16$) or adjectives ($r = –.06$; $r_E = –.07$). At the lexical level, the relationship between relative meaning frequency and word frequency was not significant ($r = –.03$; $r_E = –.11$), but the correlation with log-transformed frequency was ($r = – .12$; $r_E = –.11$), further supporting the results of

the original study, indicating that the relationship between word frequency and relative meaning frequency is, at best, very weak. The correlation with orthographic Levenshtein distance was not significant (r = –.03; r_E = .14); however, the correlation with length, in letters, was marginal but in the opposite direction as in English (r = –.10; r_E = .10). At the sublexical level, a significant correlation was observed with the number of phonemes (r = –.11; r_E = .04), but no correlation with the number of syllables (r = –.07; r_E = .08), in a word, nor with the positional bigram frequency (r = –.03; r_E = .09).

Taken together, the overall pattern of effects in Spanish is qualitatively quite similar to that in English, with only slight variation in the magnitudes of some effects, the absence of a raw frequency effect observed in the original study, and the detection of weak effects of neighborhood size and number of phonemes that were not significant in English. This general pattern of relationships supports (although not definitively) the notion that both languages have similar relationships between semantic properties, such as relative meaning frequency, and other psycholinguistic variables. Consequently, the differing degrees of skewness in the relative meaning frequency distributions across languages may, in part, be shaped by population biases in the absolute values they assign to what are—in abstract, objective terms—identical relative meaning frequencies. The results also point to the need to ensure that ambiguous words used to contrast ambiguity effects across languages are carefully matched on these metrics.

External validity

Does the dictionary's ordering of definitions agree with participants' rankings of dominant versus subordinate meanings? Studies of context-sensitive ambiguous word comprehension, in particular, require the identification of the dominant and subordinate meanings of a word (e.g., Frazier & Rayner, 1990; Klepousniotou, 2002). Having established that the vast majority of the relative meaning frequency data are loaded onto the two most frequent meanings and that most of the words effectively have two meanings, it is therefore possible to ask whether the first definition listed in the dictionary is actually the dominant meaning of that word, as determined by lexicographers. This was assessed using a sign test to determine whether the first entry in the dictionary was also the highest-frequency meaning according to the participants in each population. The results indicated that the dictionary did correctly rank order the items above chance (Europe: sign test = .67, SE = .02, p < .0001; Rioplatense: sign test = .68, SE = .02, p < .0001; df = 577). However, this rank ordering was far from perfect (expected sign test value = 1.0, vs. .5 for agreement at chance levels). Thus, there is clear value in using subjective ratings to identify the dominant and subordinate meanings of the words used in psycholinguistic experiments.

Comparison to other relative meaning frequency norms Comparing the data collected in the present study to those collected in other studies using other methods and participants from other regions of Spain can provide additional insight into dialectal differences and the validity and reliability of different norming approaches.

To the best of our knowledge, the only available dominance norms for homonyms in Spanish have come from Estévez (1991), Nievas and Cañas (1993), and Gómez-Veiga, López Carriedo, Rucián Gallego, and Cháves Vila (2010). Unfortunately, the smaller number of items and relatively low overlap between the Spanish eDom norms and these previous studies (n = 27 for Gómez-Veiga et al., n = 28 for the two other studies), as well as between each of the individual studies (ns ranging from 18 to 24) prevents drawing a very strong a set of conclusions; nevertheless, these comparisons still provide some interesting—although more tentative—insights.

Before turning to the results of the comparisons themselves, it is worth reviewing three main factors that can help frame their interpretation: (1) the norming methods used in the study, which past work has shown can influence the similarity of the resulting relative frequency estimates and their validity in terms predicting performance in other tasks (Armstrong, Tokowicz, & Plaut, 2012; Twilley et al., 1994); (2) the geographic location and associated dialectal variation that could shape the similarity between different sets of norms, as established earlier in the present work; and (3) the age of the study, because relative meaning frequency estimates have been reported to vary substantially across even relatively brief (<10-year) time intervals (Swinney, 1979; Twilley et al., 1994). A lower limit for the expected similarity can be derived from the original eDom study, wherein ratings obtained in the eastern United States in 2012 were compared to relative meaning frequency estimates obtained in western Canada in 1994 that were derived from the classification of free associates (Twilley et al., 1994). Those results showed the weakest correlation amongst all of those tested (r = .27). An initial upper limit on the expected similarity can be obtained from the Twilley et al. study, which reported a correlation between free-association-based relative meaning frequencies and measures based on sentence completion and other tasks of at least .72.

Surprisingly, despite the number of different norms tested, the range of geographic/dialectal variations in the Spanish norms (two of the studies were conducted within a single region—Granada or the Canary Islands—although the Gómez-Veiga et al., 2010, report merged data from Galicia, Andalucía, and Castilla–La Mancha), the range of different time periods that elapsed between the collection of different sets of norms (0 to 25 years), and the range of similarities in the methods employed (listing definitions vs. free association vs. rating dictionary definitions), none of the correlations between either the European or Rioplatense Spanish norms

reached significance (smallest $p = .16$; coefficients with the European relative meaning frequency ratings and each of the other sets of norms: Estévez $r = –.03$, Gómez-Veiga et al. $r = .00$, Nievas $r = .23$; coefficients with the Rioplatense Spanish relative meaning frequency ratings: Estévez $r = .23$, Gómez-Veiga $r = .16$, Nievas $r = –.15$). Even if the significance level is relaxed, the results of the Gómez-Veiga et al. and Estévez studies, both of which would be expected to correlate equally or more strongly with the European data, given the greater similarity between different European Spanish dialects, showed exactly the opposite trend. Only the Nievas data produced even numerically concordant results with the dialect differences observed here. Moreover, even if one assumes, as was observed in the original eDom study, that norms obtained using the present method are particularly unlikely to correlate strongly with past measures, the correlation within the three previous studies is still also surprisingly low, relative to the value of .72 obtained in the Twilley et al. (1994) study—only the correlation between Estévez (1991) and Nievas and Cañas (1993) reached significance and was moderately strong [$r(23) = .50$, $p = .01$; Gómez-Veiga vs. Estévez, $r(24) = .09$, $p = .67$; Gómez-Veiga vs. Nievas, $r(18) = –.31$, $p = .18$].

Collectively, and given the high correlations obtained between free associations and definition lists in the Twilley et al. (1994) study, these results reinforce previous proposals that relative frequency ratings change quite rapidly over time (e.g., Swinney, 1979). They are also consistent with the argument that these data are substantially influenced by dialectal and regional differences. This in and of itself further suggests that the collection of updated relative frequency norms should play an important part in any study involving homonyms.

Conclusion

Relative meaning frequency is a critical factor to consider in studies of semantic ambiguity. In the original eDom study, Armstrong, Tokowicz, and Plaut (2012) established that the eDom method based on norming dictionary definitions was an efficient means for producing relative meaning frequency estimates for English homonyms, and that the resulting norms showed greater external validity in predicting performance in a range of experiments. Here, we extended that work to two dialects of Spanish. The results showed that the two dialects differ considerably in terms of the relative meaning frequencies of their constituent homonyms, and the comparisons to other relative meaning frequency norms hint that these ratings may change considerably across time, as well. This clearly highlights the need for localized, up-to-date norms in order to design powerful studies of semantic ambiguity, and suggests that dialectal differences may be responsible for some discrepant effects in English. The results also suggest that the distributions of ratings may differ across English and Spanish, which, if not controlled for in experimental designs, could lead to further discrepancies in cross-linguistic studies. In quantifying the reliability of the norms, it was also established that as few as seven ratings were needed to converge on a highly stable set of ratings. Additionally, researchers can be flexible in terms of whether these ratings are collected in longer sessions with fewer participants or shorter sessions with many participants. The eDom approach is therefore very practical and requires an order-of-magnitude fewer data than other methods, such as those based on the classification of free associates. With these norms in hand, new possibilities for careful experiments studying semantic ambiguity within and across two of the most widely spoken languages are opened, which will further contribute to the growing body of work studying the commonalities and differences amongst populations who speak one or more of these languages.

Author note B.C.A. was supported by a Marie Curie International Incoming Fellowship (IIF) (No. PIIF-GA-2013-689 627784). C.Z., A.C., and J.V.L. have been supported by CSIC-UDELAR, and CZ was supported by ANII. We thank the research assistants at the BCBL and Emilia Fló in Uruguay for their assistance with data acquisition and Manuel Perea for discussion of this project.

References

Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 273–278). Austin, TX: Cognitive Science Society.

Armstrong, B. C., & Plaut, D. C. (2011). Inducing homonymy effects via stimulus quality and (not) nonword difficulty: Implications for models of semantic ambiguity and word recognition. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Expanding the space of cognitive science: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 2223–2228). Austin, TX: Cognitive Science Society.

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012a). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods, 44,* 1015–1027. doi:10.3758/s13428-012-0199-8

Armstrong, B. C., Watson, C. E., & Plaut, D. C. (2012b). SOS! An algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods, 44,* 675–705. doi:10.3758/s13428-011-0182-9

Azuma, T., & Van Orden, G. C. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language, 36,* 484–504. doi:10.1006/jmla.1997.2502

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research, 24,* 57–65. doi:10.1016/j.cogbrainres.2004.12.006

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. doi:10.3758/BRM.41.4.977

Carreiras, M., Armstrong, B. C., Perea, M., & Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends in Cognitive Sciences, 18,* 90–98. doi:10.1016/j.tics.2013.11.005

Degani, T., & Tokowicz, N. (2010). Semantic ambiguity within and across languages: An integrative review. *Quarterly Journal of Experimental Psychology, 63,* 1266–1303. doi:10.1080/17470210903377372

Degani, T., & Tokowicz, N. (2013). Cross-language influences: Translation status affects intraword sense relatedness. *Memory & Cognition, 41,* 1046–1064. doi:10.3758/s13421-013-0322-9

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods, 45,* 1246–1258. doi:10.3758/s13428-013-0326-1

Duñabeitia, J., Casaponsa, A., Dimitropoulou, M., Martí, A., Larraza, S., & Carreiras, M. (2014). *BaSp: A Basque–Spanish database*. Manuscript in preparation.

Estévez, A. (1991). Estudio normativo sobre ambigüedad en castellano. *Cognitiva, 3,* 237–271.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language, 29,* 181–200.

Frost, R. (2012). A universal approach to modeling visual word recognition and reading: Not only possible, but also inevitable. *The Behavioral and Brain Sciences, 35,* 310–329.

Frost, R., Armstrong, B. C., Seigelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences, 19,* 117–125. doi:10.1016/j.tics.2014.12.010

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A multilingual naming test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition, 15,* 594–615. doi:10.1017/S1366728911000332

Gómez Veiga, I., López Carriedo, N., Rucián Gallego, M., & Cháves Vila, J. O. (2010). Estudio normativo de ambigüedad léxica en castellano, en niños y en adultos. *Psicológica, 31,* 25–47.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61,* 29–48.

Hino, Y., Kusunose, Y., & Lupker, S. J. (2010). The relatedness-of-meaning effect for ambiguous words in lexical-decision tasks: When does relatedness matter? *Canadian Journal of Experimental Psychology, 64,* 180–196. doi:10.1037/a0020475

Hino, Y., Pexman, P., & Lupker, S. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *Journal of Memory and Language, 55,* 247–273.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44,* 287–304. doi:10.3758/s13428-011-0118-4

Klein, D., & Murphy, G. (2001). The representation of polysemous words. *Journal of Memory and Language, 45,* 259–282.

Klein, D., & Murphy, G. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language, 47,* 548–570.

Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language, 81,* 205–223.

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language, 123,* 11–21.

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1534–1543. doi:10.1037/a0013012

Lerner, I., Armstrong, B. C., & Frost, R. (2014). What can we learn from learning models about sensitivity to letter-order in visual word recognition? *Journal of Memory and Language, 77,* 40–58. doi:10.1016/j.jml.2014.09.002

Mirman, D., Strauss, T., Dixon, J., & Magnuson, J. (2010). Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science, 34,* 161–173.

Nievas, F., & Cañas, J. J. (1993). Asociados de una base de homógrafos. *Psicológica, 14,* 269–279.

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition, 28,* 657–666. doi:10.3758/BF03201255

Real Academia Española. (2001). *Diccionario de la lengua española* (22nd ed.). Madrid, Spain.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46,* 245–266. doi:10.1006/jmla.2001.2810

Seidenberg, M., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568. doi:10.1037/0033-295X.96.4.523

Seidenberg, M., Tanenhaus, M., Leiman, J., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology, 14,* 489–537.

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin, 134,* 584–615. doi:10.1037/0033-2909.134.4.584

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18,* 645–659.

Tabossi, P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language, 27,* 324–340.

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition, 22,* 111–126. doi:10.3758/BF03202766