Associative strength and semantic activation in the mental lexicon: evidence from continued word associations

Simon De Deyne (simon.dedeyne@ppw.kuleuven.be)[a,b]
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)[b]
Gert Storms (gert.storms@ppw.kuleuven.be)[a]

[a] University of Leuven, Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium
[b] University of Adelaide, School of Psychology, 5005 Adelaide, Australia

Abstract

In a word association task, the probability of producing a certain response to a cue is considered to be a direct measure of associative strength between words in the mental lexicon. The common single word association procedure is limited, since the number of words connected to a cue might be underestimated when a single response is asked. The continued association task overcomes this limitation by asking a person to generate multiple associative responses. To test whether continued strengths allow a better approximation of our lexicon, an experiment was conducted in which participants judged the associative strength between words.

Our results show that in contrast to other semantic tasks, continued strength predicts weak to moderate judgments only. Two explanations based on the sampling of information and differential semantic activation of later responses in continued association are proposed. Theoretical implications for semantic activation and methodological implications for derivation of strength are discussed.
Keywords: associative strength, semantic relatedness; word associations.

The free word association task has been used extensively to investigate processes and structure in semantic and episodic memory. The task is attractive because it is unconstrained and straightforward, and no apriori restrictions are formulated about what types of relationships between words are deemed relevant. It leads to a rich and varied source of information. Compared to constrained tasks such as feature generation, it tends to provide more thematic relations like DOCTOR - NURSE. There is increasing agreement that this thematic information determines much of how natural language concepts are used both in daily life and in language phenomena studied in the lab including semantic priming, metaphor comprehension, categorization and induction (e.g. Lin & Murphy, 2001; Wisniewski & Bassok, 1999).

An influential metaphor for the representation of this knowledge presents the mental lexicon as a weighted graph, where the structure of the links between the nodes (words) determines how words relate to each other and get their meaning. Obviously, the value of such a representation hinges on how the words are connected and on what determines the strength of these connections. The key assumption underlying the word association task, is that the number of people that generate a specific response to a cue is an indication of the strength between cue and response. Approximating the relations in the lexicon through word associations explains numerous phenomena: facilitation of word processing in associative priming (Hutchison, 2003), the probability of recall in cued-recall tasks (Nelson, Zhang, & McKinney, 2001), reaction times in lexical decision (De Deyne, Navarro, & Storms, 2012) and generation frequencies in fluency tasks (Griffiths, Steyvers, & Tenenbaum, 2007). Moreover, the overlap of the distributions of these strengths for two words indicates how semantically related they are and this is the basis of the success of lexico-semantic models such as Latent Semantic Analysis (LSA, Landauer & Dumais, 1997) and topic models (Griffiths et al., 2007).

Associative strength is central to how we process the meaning of words, but the traditional way of measuring it, through asking a participant a single word association, is not without limitations. The response frequencies from the single word association task are considered reliable only for the very strong associates, since weaker responses are often missing (Nelson, McEvoy, & Dennis, 2000). This lack of weak associations is seen as a general drawback of the word association procedure (Aitchison, 2003, p. 101) and has been responsible for questioning the results of previous findings in numerous tasks such as mediated priming (e.g., Chwilla, Kolk, & Mulder, 2000). Presumably, this reflects dominance effects where for a cue like UMBRELLA a single strong associate such as RAIN accounts for almost all responses (Nelson & Bajo, 1985). While the exact causes of dominance effects are not well understood, it is obvious that they make the response distributions overly sparse, and bias all kinds of association derived strength measures.

Recently, a large-scale continued word association database was completed involving over 70,000 participants and 3 million responses (De Deyne, Navarro, & Storms, 2012). In contrast to previous studies, a continued word association task was used in which subjects were presented a short list of stimulus or cue words and asked to give three different responses to each of these cues. The goal of the present study is to investigate how word association frequencies in continued tasks map onto associative strength. If single word associations tend to underestimate or be unreliable for weaker responses, then we would expect that using information encoded in later responses might alleviate this problem. This would support previous findings where semantic relatedness derived from continued association norms results in a better predictor of semantic tasks including pair-wise similarity judgments (De Deyne, Peirsman, & Storms, 2009), prototypicality judgments (De Deyne, Voorspoels, Verheyen, Navarro,

& Storms, 2011), and response times in the lexical decision task (De Deyne et al., 2012).

Since continued word association data only became available recently (cfr. De Deyne et al., 2012), few have studied strength derived from multiple responses and how it relates to other measures of associative strength. For instance in the study of Garskof (1965) calculated strength using continued associations to 20 cues and found that a weighted sum depending on the rank of the response correlated higher than a measure of strength that did not take into account response position. In contrast to previous work, this study presents a systematic comparison of measures of strength by looking at the contribution of continued responses alongside that of single responses using a recently proposed task in which participants judge associative strength of word pairs directly (Koriat, 2008; Maki, 2007) and compare it to single associate strength measures.

Sometimes the best way to understand a phenomenon is to take a step back. To aid the interpretation of the pattern of results from the judgment of associative strength task, the second part of this paper describes additional evidence by comparing expected strengths of continued responses with the observed strengths of these responses in the continued task. This analysis allows us to interpret quantitative differences (due to the sampling regime in continued association), and qualitative differences in terms of the types of semantic information activated in later responses.

Judgment of Associative Strength Experiment

In a series of experiments on associative strength, Maki (2007) asked subjects to estimate how many people out of 100 would consider two words to be associated. Using a similar judgment of association strength task, our goal was to find out whether continued responses provide a better approximation compared to a single response procedure.

To test this hypothesis we compared various models, starting with a simple one that predicts judgments using the word association counts of the first three response positions (R_1, R_2, R_3). Strength can be forward strength (FS), or the probability that a certain response is generated given a cue or backward strength (BS): the probability of a certain cue given a specific response. These measures are easily derived by dividing the frequency of a certain response by the total number of responses for that cue.

Method

Participants Fifty native Dutch speaking psychology students participated in exchange for course credit.

Stimuli and Materials The stimuli were selected from a set of more than 12,000 Dutch cues that were part of a large scale continued word association database described in De Deyne et al. (2012). Similar to De Deyne et al. (2012) single and multiple response strength were derived from the graph G_1 based on the first response G_2 based on the secondary and G_3 for tertiary responses. These graphs were obtained by converting the bimodal cue by response matrix to a unimodal cue by cue matrix by retaining those responses present in the set of cues. This makes it possible to get estimates of both backward and forward strength since all responses are also present as a cue in such a graph. For each cue, FS was calculated using only the first response (G_1) or including the sum of all three responses G_{123}. The cues were determined randomly subject to following conditions. Only responses that were present both in G_1 and G_{123} were considered. The difference in response strength was calculated and responses were selected that differentiated between both graphs.

A total of 80 associated cues and responses were chosen to cover the entire range of forward and backward strength between 0 and 1. All words in the judgment tasks were unique and only Dutch words were admitted that had a word frequency larger than one in the SUBTLEX-NL word frequency norms (Keuleers, Brysbaert, & New, 2010). Similar as in Maki (2007), 20 unrelated pairs such as RAFT–LION or TASK–SIN were added to the 80 related pairs. Since these do not share any associations, their forward and backward strengths equaled zero.[1]

Procedure Participants were tested during a collective session in a computer room using an online survey. Similar to Maki (2007) the subjects were familiarized with the word association task. Each participant was asked to give three responses to a set of 15 cues in a task identical to the one described in De Deyne et al. (2012). Upon completion of the word association study they were directed to the instruction page for the judgment of associative strength study and asked to estimate how many out of a hundred persons from Belgium, would give a certain association. An example was shown for a highly related pair (CAPTAIN–BOAT) and a weakly related pair (CAPTAIN–HAT). Finally, they were told to use a sliding scale to indicate their judgments and to consider the entire range of the scale from 0 to 100. A total of 100 items were presented in a randomized order and had the following format: *In a word association task, the word* X *was presented. How many people out of 100 responded with the word* Y*?* The judgment of associative strength task took about 10 minutes on average to complete.

Results and Discussion

The average of all ratings was calculated and the Spearman-Brown formula for split-half reliability was applied on the data from 50 subjects. The result showed that the ratings were highly reliable: $r_{splithalf}(100) = .99$.

The judgment of association strength as a function of normed association based on single response strength FS_1 is plotted in Figure 1. This Figure shows that weak and moderately strong normed associates are overestimated in the judgments of strength (as indicated by their relative position toward the diagonal), while strong associates tend to be underestimated. This is in line with the previous findings reported

[1] A full list of the stimuli is available from http://www.smallworldofwords.com/experiments/

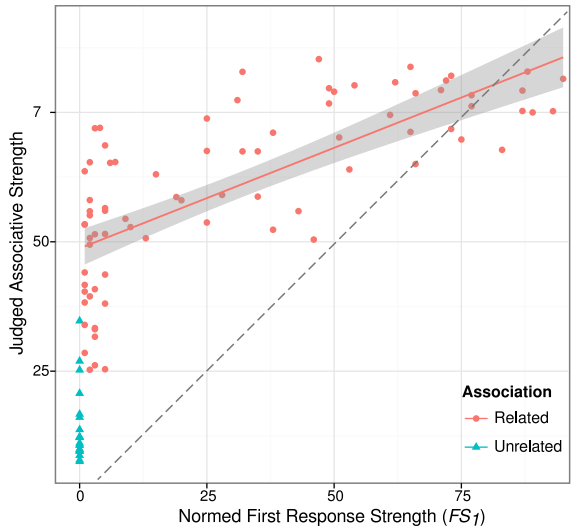

Figure 1: Scatter plot for judged and normed FS_1 together with regression line and confidence bounds.

by Maki (2007). The relative contribution of different instantiations of associative strength measures based on continued association was investigated through a series of regression analyses where we focused on straightforward predictors that corresponded to interpretable and theoretically interesting aspects of strength.[2]

Strength in related pairs In a first series of analyses, only the related items are considered as these data have non-zero values for both forward and backward strength measures. The results of the analyses are presented in Table 1. The simplest model predicts judgments of associative strength by the normed strength of the first response (FS_1). Model 1 accounts already for 56% of the variance and found a significant effect of FS_1, $\beta = .75, p < .01$. A second model is one where strength is averaged over all three responses: $FS_{123} = (FS_1 + FS_2 + FS_3)/3)$. This predictor was significant ($\beta = .70, p < .01$), but the model only captured 49% of the variance. In contrast to previous studies (cfr. De Deyne et al., 2011, 2012), the added information from R_2 and R_3 does not improve the prediction of the judgments of associations strength. Model 3 considers the possibility that judged strength is a function of both forward and backward strength of R_1. Both $FS_1 (\beta = .70, p < .01)$ and $BS_1 (\beta = .21, p < .01)$ were significant predictors and provided the best account of the data so far.

Next, we investigated if R_2 and R_3 responses provide additional information beyond that captured in R_1. Model 4 expands Model 1 by including FS_2, resulting in significant effects for FS_1 ($\beta = .75, p < .01$) but not FS_2 ($\beta = .11, ns$) resulting in little extra variance accounted for (see Table 1). Similarly, no effect was found for FS_3 in any additional analysis that was not accounted for by either FS_1 or FS_2. So these will not be discussed further.

[2]To reduce the skew in the count-based strength measures a log-transformation was used.

Table 1: Regression models (#M) for the prediction of judged associative strength. Only significant models are reported and adjusted R^2s are used throughout.

M	F-test	Regression Equation	R^2
		Related	
1	$F(1,78) = 99.8$	$69 + 22FS_1$	.556
2	$F(1,78) = 75.9$	$88 + 14FS_{123}$	.487
3	$F(2,77) = 58.9$	$83 + 21FS_1 + 16BS_1$	.594
4	$F(2,77) = 51.7$	$74 + 22FS_1 + 7FS_2$	.562
5	$F(3,76) = 52.0$	$90 + 17FS_1 + 8BS_1 + 29Rel$	.659
		25% Quantile	
1	*ns*	–	–
2	$F(1,37) = 6.4$	$71 + 91FS_{123}$	.125
3	$F(1,37) = 4.3$	$59 + 14FS_2$	.080
4	$F(2,36) = 8.5$	$102 + 16FS_2 + 43BS_1$	.284
5	$F(3,35) = 11.3$	$107 + 17FS_2 + 22BS_1 + 38Rel$	.448

A final model considered the role of relatedness. It is quite possible that when faced with uncertainty about exact strength, participants use the semantic relatedness between the cue and target to infer how strongly associated they are. Semantic relatedness was calculated as the cosine between the cue and response vector (see De Deyne et al. (2012) for additional details). Intuitively a high cosine indicates many shared associates between two words, while a low cosine indicates few shared associates. Model 5 gave the best fit of the data ($R^2 = .66$), with significant effects for both FS_1 ($\beta = .59, p < .01$) and relatedness ($Rel, \beta = .31, p < .01$). BS_1 was no longer significant ($\beta = .10, ns$).

Modeling weak strengths Still, it might be too early to conclude that normed strength from later responses never predicts strength judgments. As can be seen from Figure 1, FS_1 at the low end of the scale does not distinguish much of the observed judged data. Possibly, strength derived from later responses results in more stable estimates for those responses that occur less frequent as R_1. At this low end of the FS_1 scale, participants might make use of richer information, corresponding to information encoded in FS_2, FS_3, backward strength, or semantic relatedness.

To investigate if the weak strengths are better captured by R_2 and R_3, a subsection of the data presented was selected by placing a cut-off at the first quartile of FS_1, as most of the remaining data were not explained by FS_1.

The same models as presented before were now used to predict these data. The results for FS_1 in Model 1 confirmed the pattern in Figure 1, as it was unable to predict any data. A significant effect for summed strength FS_{123} ($\beta = .38, p < .05$) was found in Model 2, explaining 13% of the variance. Since FS_1 did not explain the data, a new model consisting of FS_2 was tested and found significant ($\beta = .32, p < .05$). The following models therefore use FS_2 rather than FS_1. In Model 4, both FS_2 ($\beta = .38, p < .01$) and BS_1 ($\beta = .47, p < .01$) were significant and accounted for 28% of the variance.

The final model including relatedness explained most of the variance (45%) with a significant effect of FS_2 ($\beta = .40, p < .01$) and relatedness ($\beta = .47, p < .01$), but no significance for backward strength BS_1 ($\beta = .24, ns$).

Together, these results support the idea that the judgment of association strength task is sensitive to normed associative strength, and closely replicates the previous findings of (Maki, 2007). However, our main goal was to investigate whether continued responses lead to better approximations of judged strength. Our findings support this hypotheses, but only for weak or moderate strengths. Since no large-scale studies have looked at the effect of continued associations, the next section will go into detail about which mechanisms might cause these results.

What factors determine the contribution of continued responses?

A question that arises from the previous findings is why strength measures that include R_2 and R_3 responses systematically improve the prediction in a variety of semantic tasks such as similarity judgment tasks or lexical decision tasks (De Deyne et al., 2012), but not in the judgment of association task. Can we provide an explanation why they capture no additional information compared to R_1 strengths at the high range of the scale?

Sampling without replacement hypothesis. A first explanation is based on the idea that continued responses are biased due to the continued nature of the task. More precisely, participants are not allowed to repeat a response. Especially when a certain R_1 association is very dominant, the proportion of participants who did not generate it as R_1 but could generate it as R_2, will be very low. Summing strengths in these cases might bias strength for such a response. In other words, the strength measures for R_2 and R_3 do not take into account this sampling without replacement. As consequence of the restriction of sampling without replacement, we expect FS_2 to be heavily biased for the strong responses, but at least capture moderate and weak strengths. If sampling without replacement is the main factor governing the observed frequencies for continued responses, then the derived expected strengths for the secondary and tertiary association response should closely agree with the observed strength for R_2 and R_3.

Given a specific cue with N different responses one can derive the expected R_2 response count for x from its probability as a first response R_1 as follows:

$$P(R_2 = x) = P(R_1 = x) \sum_{i=1, i \neq x}^{N} \frac{P(R_1 = i)}{1 - P(R_1 = i)} \qquad (1)$$

The same principle holds for the derivation of the joint expected response for R_3. For each of the 12,428 cues in G_1, the expected R_2 strengths were calculated using Equation 1. If differences between expected and observed FS_2 strengths are primarily caused by the sampling without replacement, then the expected and observed values should be similar up to some random noise. For each of the cues, the correlations between expected and observed strength distributions were obtained and had an average correlation of $r(12428) = .71 (SD = .13)$. At this point, it is not clear what determines high or low agreement. A corollary from the strength without replacement explanation is that the degree of bias in FS_2 or FS_3 will depends on the set-size or heterogeneity of the $R1$ response distribution which can be formalized as entropy H:

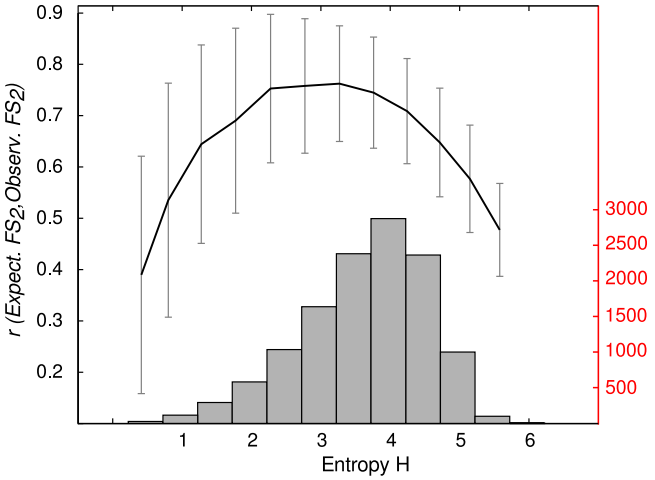

Figure 2: Averages and *SD* for correlation between expected and observed FS_2 (left-hand y-axis) grouped by entropy (H). A histogram of entropy for each cue with counts was added (right-hand y-axis).

$$H = \sum_{i=1}^{N} p_i log_2 \frac{1}{p_i} \qquad (2)$$

where N is the size of the vocabulary or number of different responses and p_i is the probability for the ith response. H increases as the responses become more heterogeneous and equals zero if all responses were identical.

Figure 2 shows the average correlations binned as function of the entropy for the cues. For cues with few responses, the correlation between expected and observed counts is lower. Similarly, the cues with a very heterogeneous response set corresponding to the high entropy words at the right-hand side of Figure 2 also exhibit lower agreement than average entropy cues. A possible explanation of the former effect is due to dominance effects previously observed in cued recall (Nelson & Bajo, 1985), where a single strong response inhibits the retrieval of other weaker ones. For these low entropy cues we expect higher utility of FS_2 or FS_3 in the judgment of associative strength assuming that the effect of dominance is removed once the response is generated and additional information becomes accessible. The latter effect could be due to unreliability, where at the high extreme cues elicit only idiosyncratic responses. Little benefit of FS_2 can be expected for high entropy cues, since there is no reason to expect very heterogeneous responses to become more coherent in the later responses. For these cues it should be dif-

ficult predicting associative strength whether this strength is based on FS_1 or FS_2. New pilot studies seem to support this entropy interaction. However, there are number of reasons why sampling restrictions cannot completely explain the observed response distributions for continued responses. First of all, this does not explain why the heterogeneity or entropy increases when more than one response per cue is asked. Second, it might be the case that for different response positions, distinct types of semantic information becomes available.

Time course of Semantic Activation Hypothesis. A possible explanation why some R_2 and R_3 responses are generated much more (or less) frequent than expected based on R_1 when sampling without replacement is taken into account stems from the idea that qualitatively different sources of information are accessed. A first possibility is that the type of response for R_2 and R_3 is influenced by the previous response beyond previously noted sampling restrictions. Such an order effect is called chaining, and can be illustrated for the cue SWISS, where MOUNTAINS is given more frequently as an R_1 (57%) than R_2 (16%), while it is expected 26% of cases in R_2. Together with the observation that SNOW is given less frequently than expected from its R_1 counts, one can assume an associative chain: SWISS $\rightarrow$ MOUNTAINS $\rightarrow$ SNOW. The presence of chaining can be quite easily investigated, and previous research suggest this phenomenon is quite rare (De Deyne & Storms, 2008).

Second, the different time course of automatic and qualitatively different types of semantic information might be a more important factor. Consider for example the cue GORILLA where MONKEY is generated in 72 times as R_1. It is expected to occur 21 times as R_2 yet occurs only 6 times. At the same time, BIG is generated 18 times as R_2, but is expected to occur 6 times at most. Perhaps linguistic or superficial information like superordinate labels precede entity properties as in this example. Both behavioral (Santos, Chaigneau, Simmons, & Barsalou, 2011) and fMRI studies (Simmons, Hamann, Harenski, Hu, & Barsalou, 2008) support the idea that gradually deeper semantic information becomes activated. For example, in an experiment by Santos et al. (2011), participants generated about 1.7 responses in a continued time delimited task. In this study, later responses tended to convey a shift from primarily linguistic responses towards taxonomic- and especially thematic- and entity-related responses.

Perhaps a better way to study the time course of semantic activation is based on a comparison between observed and expected response frequencies for continued responses given the response distribution of the first response. Such a comparison is more accurate compared to previous approaches since it is not biased by (lack of) opportunity to generate a previous response in the continued procedure. To investigate what type of information is different in the second and third response of the word associations, we calculated the expected response frequencies for R_2 and R_3 (cfr. Equation 1) and compared them with the observed response frequencies by subtracting observed from expected R_2 and R_3. A positive value indicates

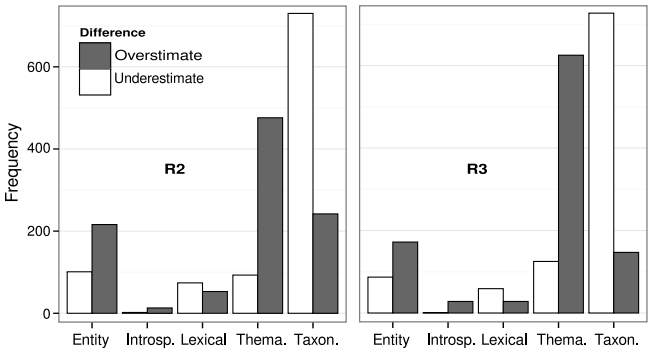

Figure 3: Distribution of semantic knowledge for observed responses in R_2 and R_3 that are either over- or underestimated based on expected R_2 and R_3 responses.

that the observed response in R_2 or R_3 is less likely to be generated than expected and this information is underestimated in R_2 or R_3. A negative value indicates that the response is generated more often than expected and is overestimated in the observed R_2 or R_3 counts. Since it is practical unfeasible to manually code all possible cue-response pairs only a subset of the data was used. For each of the +12,000 cues the most extreme (one positive, one negative) responses were listed, once for R_2 and once for R_3. Both sets were sorted and only the 1,000 most negative and 1,000 most positive differences were retained for further analysis.

The relationship between 2,000 R_2 and 2,000 R_3 cue-response pairs was coded as either as *entity*, *introspective*, *lexical*, *thematic* or *taxonomic* using similar guidelines as those described in De Deyne and Storms (2008) and Santos et al. (2011). Entity responses encode properties of the cue (e.g., MOON-YELLOW), introspective pairs encode evaluation or affect towards the cue (MOON-PRETTY), lexical attributes encode linguistic properties such as word compound completions, idioms, or rhyme (MOON-walk), thematic information could refer to agents, time and place of an action etc. (MOON-ASTRONAUT), taxonomic encodes super-,sub- and coordinates, synonyms and antonyms (MOON-PLANET). A detailed discussion of the implications for various types of semantic is beyond the scope of our illustration. For current purposes, we are mainly interested in identifying potential systematicity in qualitative response changes as a function of response position. The results in Figure 3 indicate that this is strongly the case. The largest effect is for taxonomic information which is much less likely to occur R_2 and R_3 than expected. To a lesser extend, there is also a shift where less lexical responses are generated as R_2 or R_3. The positive shift shows that entity and thematic responses are generated more frequent than expected for R_2 and R_3. These findings support the previous conclusions that linguistic information (encoded lexical) precedes conceptual types of information such as entity and thematic information. In contrast to the findings of Santos et al. (2011), our findings also show that taxonomic information is available early in the generation process.

This result also offer a potential qualitative interpretation of the contrast between a lack of effect of FS_2 in the judged associative strength task and its significant contribution in similarity judgment and other more semantic tasks if semantic knowledge related to entity features and thematic roles is better encoded in R_2 and R_3. Clearly, follow-up studies are needed to further evaluate these hypotheses.

Discussion and Conclusion

Using a judgment of associative strength task, we investigated the role of normed strength derived continued word associations. In contrast to previous reports where denser representations derived from second and third responses provided better estimates of distributional relatedness and lexical centrality (De Deyne et al., 2012), we found that the contribution of these responses is limited to weak or moderate response strengths. Moreover, in contrast to previous studies, simply summing response frequencies systematically resulted in inferior predictions for judgments of associative strength. Our interpretation of this finding is based on the notion that later responses are likely to underestimate the highest strengths due to sampling without replacement.

When comparing expected strengths under sampling without replacement against the observed strengths, the differences in R_2 and R_3 are very systematic and point out how semantic activation of types of knowledge changes over time, an issue which has been notoriously difficult to measure using other paradigms including priming. Importantly, using expected response frequencies for continued responses in comparison with actual observed response frequencies might provide a less biased baseline for tracking the time-course of semantic activation through continued association tasks. While different semantic information in continued responses strongly reflects the divergence between expected and observed counts for R_2 and R_3, it should be noted that other factors might also play a role. Since none of the responses in the association data is stemmed, it is quite likely that some part of the discrepancies will disappear when the data is processed this way. Our findings also result in a number of methodological recommendations as we have shown that ignoring sampling without replacement is problematic for low entropy cues and the use of single or combined strength measure depends on the type of task under consideration (ranging from associative to more semantic in nature).

At a theoretical level, our results challenge the main conclusions about the supposed overestimation bias of weak and moderate associates in judgments of associative strength (Maki, 2007; Koriat, 2008). The previous interpretation rests on the assumption that word association frequencies veridically reflect strength and only a small number of different responses are available (as is the case in single word association). Instead, we propose that this bias might not be due to the judgments themselves but could equally be an artifact of the single association procedure which underestimates low to medium responses.

Acknowledgments

This work was supported by a research grant funded by the Research Foundation - Flanders (FWO) to the first author and by the interdisciplinary research project IDO/07/002. Special thanks to Toon Van Borm, Steven Verheyen and Amy Perfors for helpful comments.

References

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon.* Wiley-Blackwell.

Chwilla, D., Kolk, H., & Mulder, G. (2000). Mediated priming in the lexical decision task: Evidence from event-related potentials and reaction time. *Journal of Memory and Language*, *42*, 314-341.

De Deyne, S., Navarro, D., & Storms, G. (2012). Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations. *Behavior Research Methods*, *Advance online publication. 10.3758/s13428-012-0260-7*.

De Deyne, S., Peirsman, Y., & Storms, G. (2009). Sources of semantic similarity. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (p. 1834-1839). Austin, TX: Cognitive Science Society.

De Deyne, S., & Storms, G. (2008). Word Associations: Network and Semantic properties. *Behavior Research Methods*, *40*, 213-231.

De Deyne, S., Voorspoels, W., Verheyen, S., Navarro, D. J., & Storms, G. (2011). Graded structure in adjective categories. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (p. 1834-1839). Austin, TX: Cognitive Science Society.

Garskof, B. E. (1965). Relation between single word association and continued association response hierarchies. *Psychological Reports*, *16*, 307-309.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211-244.

Hutchison, K. (2003). Is semantic priming due to association strength or feature overlap? *Psychonomic Bulletin and Review*, *10*, 785-813.

Keuleers, E., Brysbaert, M., & New, B. (2010). Subtlex-nl: A new measure for dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650.

Koriat, A. (2008). Alleviating information of conditional predictions. *Organizational Behavior and Human Decision Processes*, *106*, 61-76.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211-240.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *1*, 3-28.

Maki, W. (2007). Judgments of associative memory. *Cognitive psychology*, *54*(4), 319–353.

Nelson, D. L., & Bajo, M. T. (1985). Prior knowledge and cued recall: Category size and dominance. *The American Journal of Psychology*, *98*, 503-517.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*, 887-899.

Nelson, D. L., Zhang, N., & McKinney, V. M. (2001). The Ties That Bind What Is Known to the Recognition of What is New. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1147-1159.

Santos, A., Chaigneau, S., Simmons, W., & Barsalou, L. (2011). Property generation reflects word association and situated simulation. *Property generation reflects word association and situated simulation. Language and Cognition,*, *3*, 83-119.

Simmons, W., Hamann, S., Harenski, C., Hu, X., & Barsalou, L. (2008). fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology - Paris*, *102*, 106-119.

Wisniewski, E. J., & Bassok, M. (1999). What Makes a Man Similar to a Tie? *Cognitive Psychology*, *39*, 208-238.