Representing the Richness of Linguistic Structure in Models of Episodic Memory

Melody Dye (meldye@indiana.edu)[1], Michael Ramscar (michael.ramscar@uni-tuebingen.de)[2], & Michael N. Jones (jonesmn@indiana.edu)[1]

[1] Department of Psychological & Brain Sciences, Indiana University, Bloomington

[2] Department of Linguistics, University of Tübingen

Abstract

The principal aim of a cognitive model is to infer the process by which the human mind acts on some select set of environmental inputs such that it produces the observed set of behavioral outputs. In this endeavor, one of the central requirements is that the input to the model be represented as faithfully and accurately as possible. However, this is often easier said than done. In the study of recognition memory, for instance, words are the environmental input of choice—yet because words vary on many different dimensions, and because the problem of quantifying this variation has long been out of reach, modelers have tended to rely on idealized, randomly generated representations of their experimental stimuli. In this paper, we introduce new resources from large-scale text mining that may improve upon this practice, illustrating a simple method for deriving feature information directly from word pools and lists.

Keywords: recognition memory; word frequency; word length; feature frequency; orthographic similarity; semantic similarity; corpus analysis; vector space models

Introduction

In cognitive modeling, there is a close interdependence between representation and process. A model consists in both a ***data structure*** (an abstract representation of environmental input), and an ***algorithm*** (the process that operates over the data to simulate behavior). The choice of structure for the underlying data is critical, as it can profoundly influence the choice of algorithm. Valid representational assumptions are of vital importance, in that they reduce the degrees of freedom available to the modeler, thereby constraining model selection.

Since the inception of memory research, psychologists have relied on verbal stimuli to study learning and forgetting (Ebbinghaus, 1885). In episodic and semantic memory, the majority of data has been—and is still—generated from experiments with word lists, and memory models are routinely assessed in terms of their ability to fit data on verbal remembering (Monsell, 1991). However, when it comes to words, the choice of data structure is complicated by the fact that words vary on a remarkable number of lexical and semantic dimensions (Baayen, Milin, & Ramscar, 2016), which may or may not contribute to how they are learned and remembered. Historically, it has been impossible to reliably quantify all these points of variation. Memory modelers have thus tended to rely on randomly generated representations, which have been carefully selected to preserve the relevant properties of the data.

While this practice has been expedient, it is no longer strictly necessary. As large-scale corpora—and the technology to mine them—have become widely available (Gilquin & Gries, 2009; Halevy, Norvig, & Pereira, 2009; Recchia & Jones, 2009), it has become not only possible, but relatively straightforward, to construct not merely plausible, but accurate representations of the stimuli used in a given experiment (Baayen, 2010). This is an important advance, as problems can arise when the selected representation does not faithfully reflect the environment. For instance, global matching models of episodic memory have considerably more difficulty reproducing behavioral data when supplied with realistic semantic representations (Johns & Jones, 2010). Improving the quality of our data structures could thus improve the quality of our process models.

Further, while the words selected for memory experiments are commonly assumed to vary randomly, in line with their selection procedure, this may not always be the case. For one, certain properties — such as semantic similarity — may be systematically skewed, and thus poorly represented by a normal distribution (Johns & Jones, 2010). For another, there may be accidental variation between the word pools used by different research groups (van Heuven et al., 2014), which could produce conflicting results. Given renewed interest in replicability in the psychological and brain sciences (Open Science Collaboration, 2015), providing a more detailed account of the stimulus properties that produce a given effect should be a principal research aim (Ramscar, 2016).

The overarching goal of this paper is to enumerate a simple technique for investigating the lexical and semantic characteristics of a specific word pool, and to discuss how this can be fruitfully applied to the interpretation of empirical results in episodic memory.

Word Frequency

Word frequency is a measure of a word's occurrence in the language, and a proxy for an individual subject's experience with that word. Frequency has long been a variable of central importance in cognitive models, as it is one of the strongest predictors of verbal processing and remembering (Baayen, Milin, & Ramscar, 2016; Balota et al., 2007). In some models, frequency is treated as a causal variable—e.g., in a model of visual word recognition, frequency might function as an internal counter, in which each occurrence of an item increments its baseline activation upward (Coltheart et al., 2001). In others models, frequency is treated as an informative correlational variable, and items of a given frequency class are assigned specific feature values (Shiffrin & Steyvers, 1997).

Setting the details aside, virtually all models incorporate frequency in one respect or another. Given the significance of frequency as an explanatory variable, its accuracy of measurement, relation to other lexical and semantic variables, and instantiation in cognitive models are all matters of some theoretical importance. Yet in spite of this, many researchers are still working with outdated measurements and methods, which are not being updated as the field advances. One particularly remarkable example of this is that the Kučera-Francis norms (1967), collected fifty years ago, are still widely used among psychologists to determine word frequency. This is the case even though they have been known for decades to be unreliable (particularly for lower frequency words), and are, on assessment, consistently the worst performing norms across an array of lexical processing tasks (Brysbaert & New, 2009). Frequency values collected today are derived from corpora orders of magnitude larger.

Another source of concern is that word frequency itself is

routinely treated as a categorical variable, rather than a continuous one, even though dichotomizing a random variable can seriously jeopardize reliability (MacCullum et al., 2002; Hemmer & Criss, 2013). Further contributing to this problem, the standard method for binning words into high and low frequency bands fails to take into account the skewed nature of the distribution. Indeed, in an analysis of several classic studies, high frequency items were found to have considerably larger standard deviations than their low frequency counterparts, and a sizeable percentage of 'low' frequency items were shown to fall at, or above, what should have been the border between the groups (van Heuven et al. 2014).

That influential psychometric tests have been predicated on such unreliable measures raises serious questions about their validity (Ramscar et al. 2014). Nevertheless, this binary division remains common in both experimental design and in modeling.

Word Frequency Effects in Recognition

One domain in which it is still commonplace to bin experimental items into high (HF) and low frequency (LF) bands is recognition memory. Models of recognition offer an illustrative test case for why representational assumptions are important to cognitive modeling, and how they might be refined with simple data mining techniques. To clarify this example, we first briefly review recognition memory as an experimental paradigm and as a modeling domain.

In tests of single item recognition, subjects study a list of words, and then at test, are asked to discriminate words encountered at study (***targets***) from non-studied words (***foils***). The difficulty of the task lies in the fact that subjects must differentiate between words seen at study and words encountered in everyday life—i.e., they must distinguish between general familiarity with the test items and familiarity that is specific to the recognition task.

Global matching models have predominated as explanatory models of recognition performance (Hintzman, 1988; Murdock, 1982; Shiffrin & Steyvers, 1997). These models are premised on the idea that item recognition depends not only on the characteristics of the item itself, but also on other items present concurrently in memory. When a specific item is presented at test, the available item and context cues form a joint probe of memory. This search process yields a match value between the test item and the contents of memory. If this value exceeds some threshold, the item is recognized as 'old'; if it fails to meet this criterion, the item is rejected as 'new'. A grounding assumption of global matching models is that studied items will have higher match values, on average, than unstudied lures. However, item recognition is rarely perfect, and much effort has been expended in identifying how interference can arise at retrieval. Noise sources are frequently categorized into two types: item noise (McClelland & Chappell, 1996; Shiffrin & Steyvers, 1997) and context noise (Dennis & Humphreys, 2010). Item noise arises from spurious feature matches with other studied items; context noises arises from interference from extra-experimental contexts in which the tested item has occurred.

Among the findings that global matching models are designed to capture, one of the hallmarks is the ***mirror effect*** for word frequency: This is the finding that when HF and LF words are present in equal numbers at study, LF items are better recognized at test, garnering both more hits and fewer false alarms (Glanzer & Adam, 1985). One way to capture this frequency effect is to assign different parameter values to HF and LF words, thereby generating different distributions of feature values, and hence, of featural similarity between items. Such a representational choice reflects the fact that words are comprised of an array of surface and semantic properties that are known to vary with frequency, and to affect processing and remembering (Landauer & Streeter, 1973; Schulman, 1967).

For example, in the Retrieving Effectively from Memory (REM) model, the parameter settings generate HF items with more common, overlapping features than LF items (Steyvers & Shiffrin, 1997). Because these features are less diagnostic, the self-match between HF targets and their own memory traces is weaker than for LF targets; because they are more common, the likelihood of a chance feature match between HF targets and HF foils will be greater. This yields the canonical lower hit-rate and higher false-alarm rate for HF items.

Representational Assumptions

Global matching models have shown considerable success in capturing the relevant empirical data, ranging from word frequency effects to differential forgetting (Clark & Gronlund, 1996). Despite these undisputed successes, there are potential drawbacks in how they represent their list items. For one, these representations commonly lump together semantic, phonemic, and orthographic features into a single, indistinguishable feature set, making it impossible to tease apart how each dimension contributes to recognition performance. For another, representations are randomly generated, rather than empirically derived.

In the influential REM model, for example, a single parameter controls the mean and variability of the distribution that item features are sampled from (Steyvers & Shiffrin, 1997). To capture qualitative differences in item similarity between word frequency bands, the parameter is adjusted separately for high and low frequency items. However, the specific parameter settings are unconstrained by the actual properties of the stimulus set. Instead, parameters are set either by convention or by best fit to the behavioral data.

Concerns have been raised with this type of practice. In particular, such flexibility leaves the resulting models open to the criticism that they could be made to fit a wide variety of results (Roberts & Pashler, 2000). Conversely, they might require significant theoretical adjustments to account for the results when supplied with a realistic representation of the list items (see Johns & Jones, 2010 for an illustration). Finally, if different experiments produce contradictory results, there is no straightforward way to trace back these differences to the characteristics of the lists.

The theoretical claims of this class of models could be strengthened by deriving the model parameters directly from the lexical and semantic characteristics of the experimental word pool, or test list. This could be accomplished in a number of ways. In the simplest case, the actual feature distribution of the stimuli could be used to determine the closest choice of parameter settings. Another option would be to generate the input representation directly from the stimuli, using either the real feature values, or adjusted feature values (which could be made more robust by incorporating noise, or various smoothing mechanisms; see e.g., Chen & Goodman, 1999). Here, we detail a simple procedure for deriving feature information for lexical items as a function of their frequency class.

Corpus Investigation

The following investigation was conducted 1) to illustrate how various lexical and semantic feature information can be derived directly from word pools and recognition lists, 2) to examine how these feature values can be expected to vary as a function of item frequency, and 3) to assess whether standard word pools mimic these differences (and each other).

Verbal Properties and Frequency Class

In the study of semantic and episodic memory, different word pools make use of somewhat different sampling procedures and controls. Thus, our first goal was to establish a neutral, independent baseline, in which words were sampled without any special consideration other than frequency.

Table 1. *The Zipf scale of word frequency*

Zipf value	*fpmw*	*Examples*
1	0.01	antifungal, bioengineering, farsighted, harelip, proofread
2	0.1	airstream, doorkeeper, neckwear, outsized, sunshade
3	1	beanstalk, cornerstone, dumpling, insatiable, perpetrator
4	10	dirt, fantasy, muffin, offensive, transition, widespread
5	100	basically, bedroom, drive, issues, period, spot, worse
6	1000	day, great, other, should, something, work, years
7	10,000	and, for, have, I, on, the, this, that, you

Figure 1: The Zipf scale is a logarithmic scale that divides the frequency spectrum into seven discrete classes (van Heuven et al. 2014).

Word Frequency Words and their frequencies were extracted from the state-of-the-art 51 million word SUBTLEXus corpus (Brysbaert & New, 2009). Frequency classes were assigned according to the ***Zipf scale***, which is calculated for an individual item as $\log_{10}$(*frequency per billion words*). The Zipf scale has a number of advantages over the typical binary division between HF and LF words, namely that it is a logarithmic scale reflecting the psychological interpretation of frequency, and its divisions are fine-grained, creating seven distinct classes rather than the traditional two (van Heuven et al., 2014). For purposes of comparison, a Zipf value of 3 or lower corresponds to LF words; 4 or higher to HF words (**Figure 1**).

Recognition Lists To create recognition lists, 10 items were selected at random (without replacement) from a given frequency bin. Half of these items were labeled ***targets***, and the other half ***foils***, replicating the standard list construction procedure. This sampling procedure was repeated until there were 1000 such lists for each frequency class.

The aim was to compare lists created in each band on four dimensions: ***word length***, ***feature frequency***, and ***orthographic*** and ***semantic similarity*** of targets to foils. These particular dimensions were chosen to be illustrative, and because they are known to be important contributing factors to item recognition. For word length and feature frequency, counts were computed for each item, and averaged over the entire list. For orthographic and semantic similarity, the similarity of each target to the distractors present at test was computed, and similarly averaged.

To preface, these analyses successfully replicate well-established findings on each of these dimensions, while providing a straightforward method for determining the actual empirical trends of a given frequency range, or item set.

Methodology Notably, the comparatively small number of types in the higher frequency ranges placed constraints on the construction of recognition lists (**Figure 2**). Specifically, list length was necessarily kept small, and while lists were created for Zipf values 1-6, 7 was excluded, as it comprised only 13 distinct word types, all of them function words.

This type distribution is a consequence of the universal scaling law for word frequencies, commonly known as ***Zipf's Law*** (1949). The idea is this: Say, an English text is selected, and each of the word types that occur in the text are arranged in order of their frequency, from most to least common, and assigned a numerical rank. Then, the full contents of the text – that is, all of its word tokens – are thrown into a bag, shook, and one word is selected at random. Zipf's Law states that the probability of drawing a given word is inversely proportional to that word's rank ordering. The law formalizes the notion that while a few words in a language are very common, the greater part are exceedingly rare.

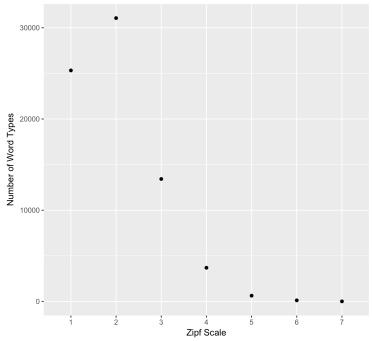

Figure 2: The number of distinct word types in the SUBTLEXus corpus for each value of the Zipf scale.

Word Length Word length, whether computed in terms of letters or phonemes, has an inverse relationship with frequency, with word lengths tending to increase as frequency declines (Piantadosi et al., 2011; Sigurd, Eeg-Olofsson, & Van Weijer, 2004; Wright, 1979; see **Figure 3**).

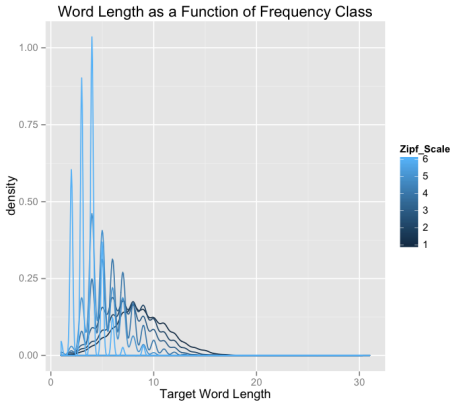

Figure 3: Average word length of list items increases as frequency declines.

Feature Frequency Feature frequencies represent the empirical n-gram frequencies of individual letters and letter combinations, and can be conceptualized as a measure of orthographic distinctiveness (**Figure 4**).

Feature frequency is known to vary with word frequency. On average, rarer words contain both more unusual letters, and more unusual combinations of letters (Malmberg et al. 2002; Zechmeister, 1969).

Orthographic similarity Orthographic similarity was computed as Levenshtein edit distance, a string metric that calculates the minimum number of edits (such as insertions, deletions, or substitutions) required to transform one word into the other (**Figure 5**).

Given that rare words are more orthographically distinctive (Landauer & Streeter, 1973; Andrews, 1992), it stands to reason that in a recognition list context, they should be less orthographically similar to frequency-matched distractors than more common words (Hall, 1979).

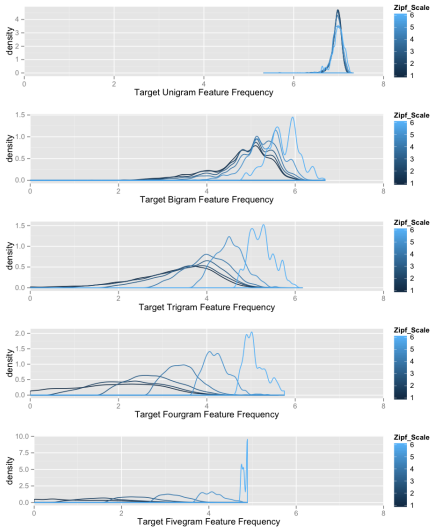

Figure 4: The five panels depict the average feature frequencies of list items in SUBTLEXus as a function of their Zipf value. The overall trend indicates that higher frequency items are comprised of higher frequency features. Moreover, the larger the n-gram, the greater the separation between frequency classes. For unigrams, a more pronounced pattern of separation between Zipf bands is observable when minimum (rather than average) feature frequency is used.

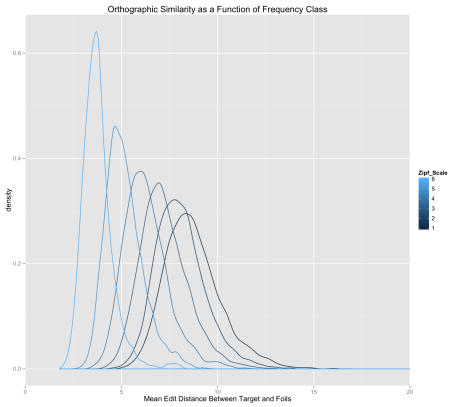

Figure 5: Average orthographic similarity between targets and distractors declines as a function of frequency.

Semantic similarity Semantic similarity values were obtained from word2vec trained on the 300 billion word Google News corpus. word2vec is a two-layer neural network that produces word embeddings (Mikolov et al., 2013), and is considered state of the art in semantic space modeling (Baroni, Dinu, & Kruszewski, 2014). word2vec was implemented with *gensim*, a Python framework for vector space modeling (Řehůřek & Sojka, 2010), which adopts the continuous skip-gram architecture. The skip-gram model weights proximate context words more highly than distant ones, yielding better results for lower frequency words.

In a recognition task in which list items are randomly sampled from a given frequency band, the semantic similarity between targets and distractors should tend to decrease with frequency (**Figure 6**). This outcome is all but assured by the distributional properties of the lexicon: In the SUBTLEXus corpus, LF words comprise 80% of word tokens (van Heuven et al., 2014) and fully 94% of word types (**Figure 2**). The semantic spread from which LF words are sampled will thus be far greater than that for HF items.

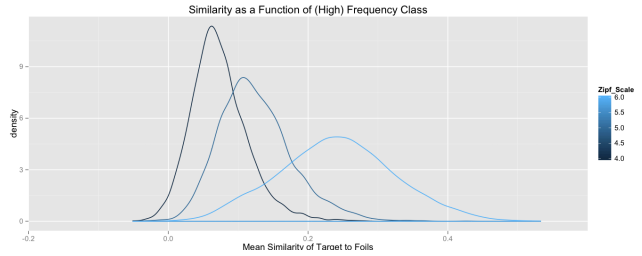

Figure 6: Average semantic similarity between targets and distractors declines across the HF range of the Zipf scale, implying that a set of randomly sampled words will be less semantically similar, on average, the lower their frequency class.

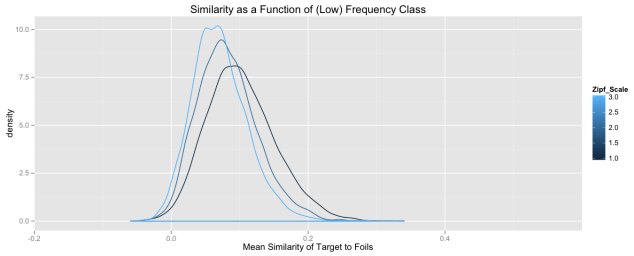

Figure 7: Average semantic similarity between targets and distractors across the LF range of the Zipf scale. While a slight (*ns*) trend in the opposite direction is observable in the lower range of the scale, this is almost certainly a methodological artifact. If the missing data in Figure 8 is included as 0-counts, the apparent trend reverses, and the pattern resembles that seen in Figure 6.

In making these calculations, there is an important methodological issue to consider—in particular, the problem, well-known to linguists, of ***data sparsity*** (Sinclair, 1997): While any given sample of language will provide ample evidence about its common words and phrases, it will provide little or none about its rarer, more informative elements (Church & Gale, 1995). Not only will many perfectly legitimate words (and word co-occurrences) fail to occur in even very large swaths of text, but even most of those that do will occur only a few times, making their estimation unreliable. This is the basic problem of data sparsity and it is one that plagues semantic similarity analyses in the lower frequency ranges (**Figures 7, 8**).

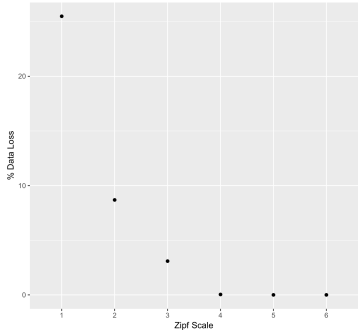

Figure 8: Data loss for the semantic similarity analyses as a function of frequency class. Semantic similarity values were not available for all the words sampled, and the proportion of words with no data points grew as frequency decreased. For Zipf rank 1, fully 25% of data was lost.

Figures 6 and **7** show the similarity distributions for item pairs that were known to our word2vec model. However, given the significant data loss for LF items, looking solely at returned values constitutes selection bias, as it implies that unobserved pairs—for which the model cannot supply a score—likely have the same distributional properties as observed pairs. In fact, it is reasonable to assume that unobserved pairs are much less similar, on average. One way of addressing this issue is to assign item pairs with null values a similarity score of 0. When these scores are included, the trend observable in the HF range (**Figure 6**) is also clearly observable in the LF range.

In the absence of knowledge, assigning 0-counts is a useful heuristic. However, given that problems with data sparsity increase as frequency declines, this solution may disproportionately penalize the lowest frequency words. In future work, similarity-based smoothing techniques might be used to better estimate similarity values for unobserved pairs (c.f. Yarlett, 2007).

Interim Summary Our analyses of words in the SUBTLEXus corpus replicates and extends a number of well-known findings on the relationship between a word's frequency and its lexical and semantic features, including that:

1) ***word length*** increases as word frequency declines,
2) ***feature frequency*** increases with word frequency, with the rate of increase dependent on feature length,
3) ***orthographic similarity*** between targets and foils increases with word frequency,
4) ***semantic similarity*** between targets and foils increases with word frequency (though the calculation of similarity scores for LF item pairs requires careful consideration).

Available Analyses In the analyses reported here, pure lists were created for each frequency class, average feature information was extracted, and similarity measures were computed as a function of the mean similarity of a target to its foils. The purpose of this was largely illustrative; many variations on this procedure are possible, depending on the requirements of the model, or the empirical task.

One obvious choice point is the sampling method. For example, word selection could be constrained by specific lexical properties (e.g., limited to nouns, or words of length *n*), as is common practice in the design of word pools. Similarly, list composition could be varied by sampling specific proportions of words from different frequency bands.

Another matter of some importance concerns the choice of comparisons and statistical measures. Similarity can be computed relative to other targets, distractors, or both; it can also be calculated as an average, or in terms of "max" similarity (e.g., the top 10% of most confusable items). Likewise, when assessing the use of rare letters and rare letter combinations, it may be more useful to know the minimum feature frequency, or the median, rather than the mean.

Finally, while we chose to delimit our focus to just a few dimensions, there are many more lexical properties that systematically vary with frequency. For instance, rare words are more likely to be judged as abstract (Galbraith & Underwood, 1973; Pavio, Yuille, & Madigan, 1968), to be acquired later (Carroll & White, 1973), and to be regular (Bybee & Hopper, 2001).

Word Pools

In the study of semantic and episodic memory, different word pools make use of somewhat different sampling procedures and controls. One concern is that different word lists may vary in systematic ways from each other, producing variability in results; another is that they may have distinctly different properties from the language 'at large'. To check the validity of these worries, we compared the word pools of two representative cognitive memory labs, with an average h-index among the principle investigators of 20, and published theoretical disagreements. These word pools were compared against a recognition word list devised by Dye, Jones, & Shiffrin (2017) (**Figures 9, 10**).

The Dye et al. (2017) word list was deliberately constructed to increase the semantic and orthographic similarity of LF items, as reflected in **Figures 9** and **10**. In a recognition list experiment, this had the predicted effect of diminishing the standard mirror effect for word frequency, by bringing the false alarm rate for low and high frequency items into line.

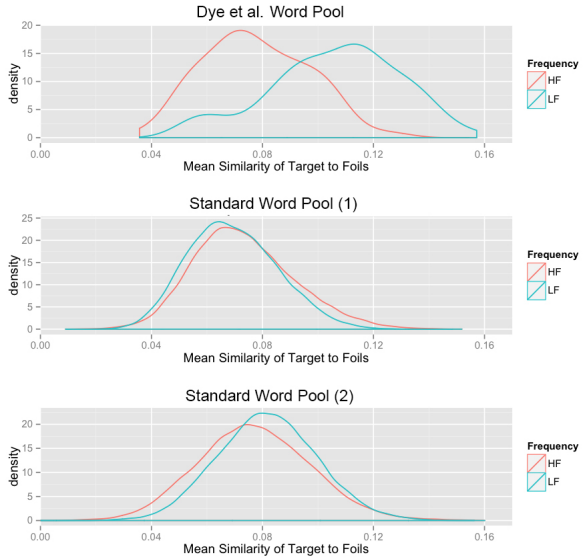

Figure 9: A comparison of average semantic similarity of targets to foils across three word pools.

Notably, while the Dye et al. word list clearly differs from the two standard word pools, these word pools are not identical to each other either. In particular, though both pools are similarly distributed in terms of frequency and semantic similarity among items, in Word Pool 2, orthographic similarity among items is substantially increased compared to Word Pool 1, and is matched across HF and LF items. This may produce differences in reported results, as orthographic similarity is known to modulate false alarm rates (Malmberg, Holden, & Shiffrin, 2004).

Finally, it is worth noting that none of these 'controlled' word pools reflect the properties expected from random sampling, as illustrated in our exploration of the SUBTLEXus corpus. In particular, while the distribution of orthographic and semantic similarity values for LF and HF items are largely overlapping for the standard word pools (**Figures 9, 10**), a truly random selection of these items shows significant separation between frequency bands (**Figures 5, 6**).

These examples illustrate how the properties of word lists can be readily and fruitfully compared both to each other, and to larger corpora. In future work, we plan to expand this analysis to include more widely used word pools, such as the Toronto word pool (Friendly, Franklin, Hoffman, & Rubin, 1982), a modified version of the Kucera & Francis word pool (1967), and a categorized word pool (Murdock, 1976).

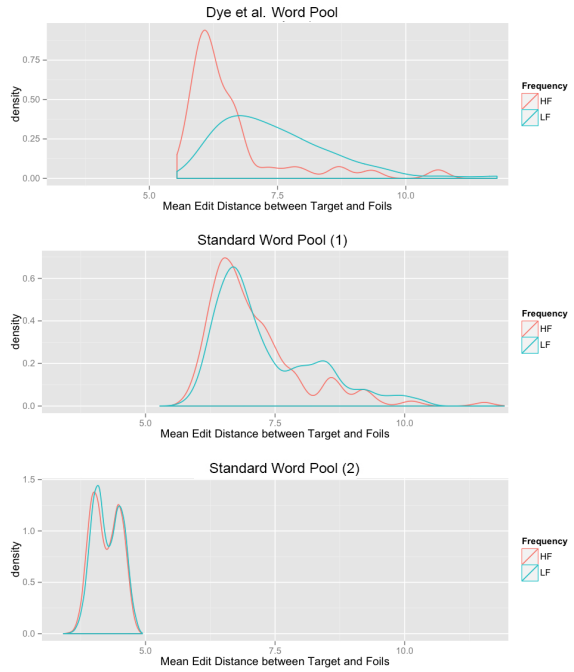

Figure 10: A comparison of average orthographic similarity of targets to foils across three word pools.

General Discussion

While work in text mining and natural language processing has considerably refined our understanding of the statistical nature of language, not all of these insights have successfully crossed over to memory research. This problem is not without remedy. In this paper, we have taken seriously the problem of furnishing an adequate description of the linguistic environment, in keeping with the roboticist Rodney Brook's famous injunction that "the world is its own best model". Analyses such as those reported here are useful in a number of different dimensions: they can be employed to deliberately control the properties of episodic word lists; they can yield a principled means for adjusting model parameter settings to reflect the properties of the specific stimulus set; and they may be useful in explaining discrepancies in published empirical results, aiding replicability. Our broader hope is that integrating more realistic representations of verbal stimuli into models of episodic memory may inform the design and interpretation of experiments and constrain the choice of process model.

Acknowledgments

This research was funded by an NSF graduate fellowship to MD. Analyses required the use of Karst, Indiana University's high-throughput computing cluster. Many thanks to Brendan Johns, Gregory Cox, and Rui Cao for insightful comments and discussion.

Notes

The density plots presented in Figures 3-7 are generated by the ggplot2 library in R, and visualize the distribution of items in each frequency class over specific dimensions of interest, using a kernel smoothing function.

References

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2), 234–254.
Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5(3), 436–461.
Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 1–47.
Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3), 445–459.
Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238–247.
Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. B*ehavior Research Methods*, 41(4), 977–990.
Carroll, J. B., & White, M. N. (1973). Age-of-acquisition norms for 220 picturable nouns. *Journal of Verbal Learning and Verbal Behavior*, 12(5), 563–576.
Chen, S. F., & Goodman, J. T. (1999). An Empirical Study of Smoothing Techniques for Language Modeling. *Computer Speech and Language,* 13, 359–394.
Church, K.W., & Gale, W.A. (1995). Inverse Document Frequency (IDF): A Measure of Deviation from Poisson (pp. 121-130). In P*roceedings of the Third Workshop on Very Large Corpora*.
Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, 3(1), 37–60.
Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204–256.
Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108(2), 452–478.
Dye, M., Jones, M., & Shiffrin, R. (2017). Vanishing the mirror effect: The influence of prior history & list composition on recognition memory. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.
Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology.* New York: Dover.
Friendly, M., Franklin, P.E., Hoffman, D., & Rubin, D.C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods and Instrumentation*, 14(4), 375-399.
Galbraith, R. C., & Underwood, B. J. (1973). Perceived frequency of concrete and abstract words. *Memory & Cognition*, 1(1), 56–60.
Gilquin, G. & Gries, S.T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistics Theory*, 5(1), 1-26.
Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2), 8–12.
Hall, J. F. (1979). Recognition as a function of word frequency. *The American Journal of Psychology*, 92(3), 497–505.
Hemmer, P., & Criss, A. H. (2013). The shape of things to come: Evaluating word frequency as a continuous variable in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(6), 1947–1952.
Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95, 528-551.
Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review,* 17(5), 662–672.
Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 12(2), 119–131.
Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4), 616–630.
Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, 30(4), 607–613.
McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory, *Psychological Review*, 105, 724–760.
Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*: arXiv:1301.3781
Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G.W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Erlbaum.
Murdock, B.B. (1976). Item and order information in short-term serial memory. *Journal of Experimental Psychology*, 105(2), 191-216.
Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89, 609-626.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716–aac4716.
Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1), 1–25.
Piantadosi, S. T., Tily, H., Gibson, E., & Kay, P. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 108(9), 3526–3529.
Ramscar, M. (2016). Learning and the replicability of priming effects. *Current Opinion in Psychology*, 12, 80-84.
Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, 41(3), 647–656.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358–367.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4, 145–166.
Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science*, 9, 211-212.
Sigurd, B. Eeg-Olofsson, M. & Van Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica*, 58(1), 37-52.
Sinclair, J.M. (1997). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora*. Longman.
van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6), 1176–1190.
Wright, C. E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory and Cognition*, 7(6), 411-419.
word2vec: https://code.google.com/archive/p/word2vec/
Yarlett, D. (2007). Language learning through similarity-based generalization. Unpublished doctoral dissertation: Stanford University.
Zechmeister, Z.B. (1969). Orthographic distinctiveness. *Journal of Verbal Learning and Verbal Behavior*, 8(6), 754-761.
Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley: Cambridge, MA.