Behavior Research Methods
2006, 38 (2), 190-195

ARTICLES FROM THE SCIP CONFERENCE

Word frequency effects in high-dimensional co-occurrence models: A new approach

CYRUS SHAOUL and CHRIS WESTBURY
University of Alberta, Edmonton, Alberta, Canada

The HAL (hyperspace analog to language) model of lexical semantics uses global word co-occurrence from a large corpus of text to calculate the distance between words in co-occurrence space. We have implemented a system called HiDEx (High Dimensional Explorer) that extends HAL in two ways: It removes unwanted influence of orthographic frequency from the measures of distance, and it finds the number of words within a certain distance of the word of interest (NCount, the number of neighbors). These two changes to the HAL model produce measures of word neighborhood density that are reliably predictive of human lexical decision reaction times.

The HAL (hyperspace analog to language) model of lexical semantics (Burgess, 1998; Burgess & Livesay, 1998; Burgess, Livesay, & Lund, 1998; Burgess & Lund, 1997, 2000; Lund & Burgess, 1996) uses the frequency of word co-occurrence to build a high-dimensional vector space. The HAL model uses the context of a word's usage to find the neighbors of a word by calculating the distance between all global co-occurrence vectors in this space. These are vectors that contain information about the co-occurrence of a word with every other word in a language. In this article, we assess the value of the HAL model in predicting human behavioral measures of lexical access, and we propose ways in which the measure may be made more relevant for this purpose. In particular, we focus on how HAL's co-occurrence distances are contaminated by the frequency of the target word, and how this contamination can be eliminated. In looking for ways to strengthen HAL's predictive relevance, we started by analyzing a previously reported measure of neighborhood density: the average distance of a word's neighbors. The measure has been called the *semantic distance* (SD; Buchanan, Burgess, & Lund, 1996; Buchanan, Westbury, & Burgess, 2001), though it is actually a neighborhood density measure, not a measure of distance. To avoid confusion, we will call this measure Burgess semantic density (BSD). We found BSD to be highly correlated with orthographic frequency (OFREQ) in a nonlinear relationship, and so we investigated ways to modify the HAL model to remove this unwanted influence of OFREQ.[1]

The second improvement that we have made is the addition of a new way to measure neighborhoods. We will propose a new way of defining what a "neighbor" is, and build new measures of neighborhood density and average neighbor proximity from this definition.

We have implemented a software system called HiDEx (High Dimensional Explorer) that reimplements and extends HAL in the ways described above. Using HiDEx, we are able to calculate measures of co-occurrence density and average neighbor proximity for words, and then use existing databases of behavioral norms to find the correlations. With this method of norm analysis, we found that our modifications to the HAL model produce measures of word neighborhood density that are correlated with lexical decision reaction times.

HAL CAN BE IMPROVED

Over time, psychological models are refined and improve. HAL is no exception. Here we will focus on two issues with HAL: the dependency of HAL on the particularities of the corpus, and the contamination of neighborhood measures by OFREQ.

Variability Across Corpora

The HAL model uses a large corpus of text as the basis of its vector space. For every corpus of text, a lexicon of words and the respective OFREQ for each word can be calculated. Each time a different corpus is used as the input to the HAL model, the occurrence frequencies and co-occurrence frequencies will differ by some amount because of the nature of the corpus (different authors, registers, or content). If these frequencies vary widely across corpora, the vectors produced by HAL may also differ greatly. For this reason, the first question that must be answered is, how much do frequencies and co-occurrence frequencies differ across different corpora of 300 million words or greater, the size of the corpus used by HAL? This answer is crucial for the long-term viability of corpus-based models of lan-

Correspondence concerning this article should be addressed to C. Shaoul, Department of Psychology, University of Alberta, P220 Biological Sciences Building, Edmonton, AB, T6G 2E9 Canada (e-mail: cyrus.shaoul@ualberta.ca).

guage. Other groups building HAL-like models have also noted that there is a need to take another look at HAL and OFREQ (Lowe, 2001; Rohde, Gonnerman, & Plaut, 2004; Song, Bruza, & Cole, 2004), and Burgess and Lund (2000) have noted that the USENET was the preferred source of text for HAL because of its variety of authors and content. Each corpus has unique properties, and in some types of research, the specific properties of a corpus may be interesting to the research questions being asked. In context-based semantic models, on the other hand, a broad, even coverage of all genres, registers, and voices is of the utmost importance.

Furthermore, the corpora used by different research groups are often not publicly available, and so all groups that attempt to replicate experiments done with HAL may be forced to use different corpora. For these studies to be considered broadly relevant, the models should produce approximately the same results with any large set of English documents. If not, the minimum size of a corpus for use by a HAL-based model might have to be increased until this condition is met. There is one other option: A core set of corpora could be assembled and offered freely to all. This would allow comparisons independent of the corpora used. We look forward to the day when such a set of corpora becomes available. Currently, we are forced to deal with the myriad of corpora available.

To find out how frequencies differ among corpora, we compared the OFREQ of the same 47,622 words in four corpora: three large corpora of recent vintage and the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995). We found that OFREQ for these words was highly variable across corpora for words in the middle and lower frequency ranges.

Table 1 shows representative correlations between the OFREQ of the words in two different corpora: the AQUAINT (Graff, 2002) corpus, a 400 million word newswire corpus, and our own MIXED corpus, a 400 million word Web-derived corpus. Words that occur under 10 times per million had very low correlations between their OFREQ in these two corpora. This means that the majority of the words in the lexicon have different OFREQ in these two corpora. We found similar results with the correlations between word frequencies in these corpora and those in the CELEX and USENET corpora.

Table 1
Comparison of Word Frequencies for Two Large Corpora of English Text (Between Our AQUAINT and MIXED Corpora)

Frequency Category of Words	OFREQ (Words/ Million)	r^2 (Correlation Between OFREQ in Both Corpora)	Percent of Words in the Lexicon
Top 100 words	$864 \le f < 53{,}000$	.93	0.2
High OFREQ	$100 \le f < 864$	.09	2.1
Med–High OFREQ	$10 \le f < 100$	.13	11.9
Med–Low OFREQ	$2 \le f < 10$	.001	20.9
Low OFREQ	$0 \le f < 2$	.23	65.2
First Quartile	$3.8 \le f < 53{,}000$	.93	25
Second Quartile	$0.85 \le f < 3.8$	.003	25
Third Quartile	$0.22 \le f < 0.85$	.01	25
Fourth Quartile	$0 \le f < 0.22$	.06	25

Co-occurrence Frequency and the BSD Measure

After concluding that the degree of variability of OFREQ was high for many words across corpora, we wanted to see whether OFREQ was also influencing the measures produced by HAL, such as BSD. Since we did not have access to the original USENET corpus used by Lund and Burgess (1996), we collected our own 284 million word USENET corpus and used the frequencies of words in this corpus for our analyses.

We received BSD measures for 4,218 words (mean OFREQ = 120 words/million, mode = 2.8 words/million, $\sigma = 548$) from Curt Burgess that were made using his 300 million word USENET corpus. We then used NUANCE (Hollis & Westbury, 2006) to characterize the relationship between BSD and OFREQ from our USENET corpus. NUANCE is a tool that searches solution space for nonlinear relationships between data sets. NUANCE found a nonlinear relationship, BSD = $1/\surd$(OFREQ), that correlated at $r = .71$, $r^2 = .5$, $p < .0001$ (see Figure 1). In essence, half of the variability in the BSD can be explained by OFREQ once it has undergone this nonlinear transformation. This suggests a bias for the BSD measure to be greater for words that occur more often.

If OFREQ is correlated with BSD, and if the OFREQ for many words is significantly different in different corpora, the corpus used to build the vectors in HAL will influence the BSD measure. We have compared frequencies from three very large corpora. The correlations of their OFREQs are low for all but the most frequent words. HAL can be improved if the issues with the influence of OFREQ can be resolved, by defining a measure that captures co-occurrence but that is less sensitive than BSD to OFREQ.

IMPROVING HAL

Our goal was to modify HAL in such a way that we would have a measure of a word's neighborhood's size and density in global co-occurrence space that was not sensitive to the OFREQ of the word. At the same time, this measure would still need to have predictive power for behavioral phenomena. First, we changed the method of vector normalization, and then we defined two new measures, both of which are relatively impervious to the influence of OFREQ. Our tool for this work is a custom software system developed by the staff of our laboratory called HiDEx (High Dimensional Explorer).[2]

HiDEx: An Implementation of the HAL Model

Collecting and preparing texts for the corpus. The first step in doing our research was to build four corpora. We used many different techniques to collect text. We used preexisting research corpora as well as our own corpora that we built from Internet sources. One source of text was

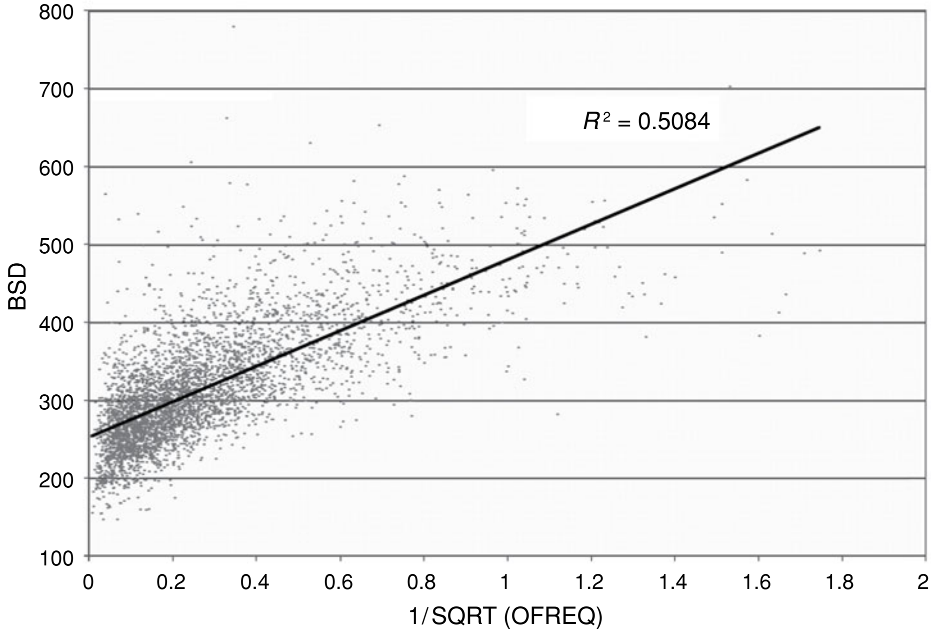

Figure 1. Plot of BSD versus 1/√(OFREQ) for 4,218 words.

the AQUAINT corpus, consisting of 375 million words of *New York Times*, Associated Press, and Xinhua wire news stories (Graff, 2002). We built one corpus with the use of Web crawling tools such as Heritrix (Mohr, Stack, Ranitovic, Avery, & Kimpton, 2004). We collected 380 million words of personal writings, and then added text from free collections, such as Project Gutenberg (post-1850), Wikipedia, and other electronic books, bringing the total for the MIXED corpus to 452 million words. We also collected 284 million words of USENET text for our USENET corpus. The fourth corpus is a concatenation of the first three corpora called the COMBINED corpus, which contains 1.1 billion words. We used these four corpora to investigate the effects of the corpus itself on the model's performance.

We processed the text with custom programs to make it usable in our model. Text has many properties that make the detection of equivalent words impossible at times. These include capitalization (Tree:tree), orthographic variation (*profit-sharing*:*profit sharing*), and other properties. The following steps were necessary to ensure reliable detection and elimination of unwanted text, as well as reliable detection of orthographic variants. We used statistical language detection techniques (Cavnar & Trenkle, 1994) to remove any non-English documents from the corpus. We then used heuristics to remove any parts of the texts that were obviously used for navigational purposes (such as *Home* or *Site Map*). We removed the HTML tags and all numbers from HTML files. Then all words were converted to uppercase letters to eliminate differences in capitalization. To avoid counting possessives as nonwords, we separated all possessive endings (*'s*) from the words to which they were attached by inserting a space (excluding contractions of *is*). Finally, all words that were hyphenated had the hyphen replaced by a space.

Next, all words not in the lexicon were replaced with the string "***" instead of being deleted from the corpus. In this way, the number of words separating two words in the lexicon could be counted, whether the intervening words were members of the lexicon or not. Since almost all words with a frequency of 2 words per million were included in our lexicon, we did not lose a significant amount of co-occurrence information in this process.

Building the co-occurrence matrix with HiDEx. The co-occurrence window size was set to be 20 words (10 on either side of the word) as it was in HAL. This window was applied to the corpus, and a count was made of all the co-occurrences of words at all positions in the window. After this was done for each word, a three-dimensional matrix of data was created. The three dimensions in this case were window position, target word, and co-occurring word. Since our lexicon contained approximately 50,000 words, the actual size of this space was around $50{,}000^2 \times 20$, or 5×10^{10} elements. Since 89% of these elements contained the number 0, we stored only the nonzero elements and their indices. Next, the co-occurrence frequencies in the window for each word were weighted. Since we wanted to compare our algorithm with the HAL results of Lund and Burgess (1996), we implemented the identical linear ramp weighting scheme, where each frequency was multiplied by a weight inversely proportional to its distance from the target (for a window size of 20, the frequency of words adjacent to the target word was multiplied by 10, the next by 9, and so on). This weighting process summed the values in each window, producing two numbers, the weighted sum for the forward window and the weighted sum for the backward window. These values then made up a new, two-dimensional global co-occurrence matrix that was 50,000 × 100,000 in size. The values in this matrix were still nonnormalized, and HiDEx normalizes the word vectors in the manner described below.

The new normalization algorithm. With the use of any type of orthographic frequency information in a calculation, there is a danger that the highly skewed distribution of word frequencies (a small number of very frequently used words, and very large number of lower frequency words) will bias the results. In particular, since high-frequency words will have much larger values in their vectors and many more nonzero elements in their vectors, they will have great impact on lower frequency words in the model. This "frequency bias" has been noted by Song et al. (2004): "Even after getting rid of the rare and stop words, however, the weighting scheme of HAL is still frequency biased—a small number of most frequent words tend to get higher weights in any HAL vector, due to their high frequency and so caused chance co-occurrence." As described by Burgess and Lund (2000), HAL divides all the elements (1 to j) in each vector by the vector length $\sqrt{(\sum w_j^2)}$, where w is the value in each element. Unfortunately, division by vector length may not remove the influence of word frequency, owing to the following confound. The vector length may not be correlated with orthographic frequency; it is actually correlated with the co-occurrence frequencies for each word. The vector length could be very different for two words with the same OFREQ if the number of words with which they co-occur varies, and if the position with which they co-occur varies. The weighting function used in the previous step gives more weight to words that co-occur close together than to those that co-occur far apart. For this reason, we felt it necessary to find a simpler, more reliable way to remove the frequency bias.

We chose to normalize all the elements in each word vector in HiDEx by dividing each element by the OFREQ of the word in the corpus instead of the vector's magnitude. The simulations in the following sections will address the effect of this change on our measures of neighborhood density.

Calculating the distances and the neighborhoods. There is ”noise“ in textual information, which comes in the form of low-frequency words that necessarily have very few occurrences. Some of these chance co-occurrences will be spurious. As in HAL, HiDEx uses only the vectors of the words with the most information, but instead of using the vectors with greatest variances, as HAL does, we use the vectors of the words with the highest OFREQs for the distance calculations. We call this parameter *context size* in HiDEx; it can be set to any number smaller than the lexicon size (typically set to 9,000 vectors). Following removal of all vectors of the words not in the context, HiDEx consists of a two-dimensional matrix of width "lexicon size" and height "2 ∗ context size." A distance metric can then be used to measure the distance between any two word vectors. In the HAL model, Euclidean distance was chosen as the metric, so we also used this metric, calculating the distance between two vectors to be

$$\sqrt{\sum_j \left(a_j - b_j\right)^2}$$

We multiplied this number by an arbitrary scaling factor of 10 to give us distances that were typically in the range of 1–300, making them easier for humans to read and understand.

Two New Measures: NCount and ARC

Neighbor Count, NCount. To further improve on HAL, we added the ability to define variable sized neighborhoods. The method used by Buchanan et al. (1996; Buchanan et al., 2001) of defining a neighborhood was to find the *N* closest words. Their measure of semantic density was defined to be the average of the distances between the target word and its *N* closest neighbors. Averaging the distances of all the neighbors can mask the distribution of those distances. Since we can make no assumptions about the distribution of the neighbor distances, averaging can produce two identical measures of density for two very different distributions. Instead of averaging *N* distances, we decided to use the standard deviation of all the interword distances to define a threshold, and measure the number of words closer than the threshold. We call this measure NCount (Neighbor Count).

BSD is a centroid, but the radius of the space that it measures varies for every word while the number of neighbors stays constant. In contrast, NCount uses a constant radius, with a variable number of neighbors, which allows us to compare the density across all word neighborhoods in an unbiased way. The method of finding NCount has three steps. First, we create the set of word pairs that have co-occurred at least once in the corpus. From this very large set, we randomly choose a subset (5%–10% of the total) of these pairs, and then calculate the distances between the words for each pair. Finally we find the standard deviation (StdDev) of all the distances. We can then set a "neighborhood membership threshold" at 1.5 times the StdDev. This cutoff point was chosen ad hoc in this initial investigation, but it may change as more is known about this measure. This new definition of the neighborhood has an interesting consequence: Some words will have no neighbors, and others will have far more than 20. NCount, the number of neighbors, is our new measure of neighborhood density.

NCount and OFREQ. We tested the correlation between NCount and OFREQ for the same 4,218 words that we used in the BSD comparison. We found that NCount was correlated at $r = .08$, $r^2 = .0065$, $p < .0001$ (see Figure 2). This correlation was much smaller than the correlation for BSD, showing that NCount is largely unpredictable from OFREQ.[3]

NCount and LDRT. To test the psycholinguistic plausibility of NCount, we performed a simulation of a lexical decision reaction time (LDRT) experiment. Using

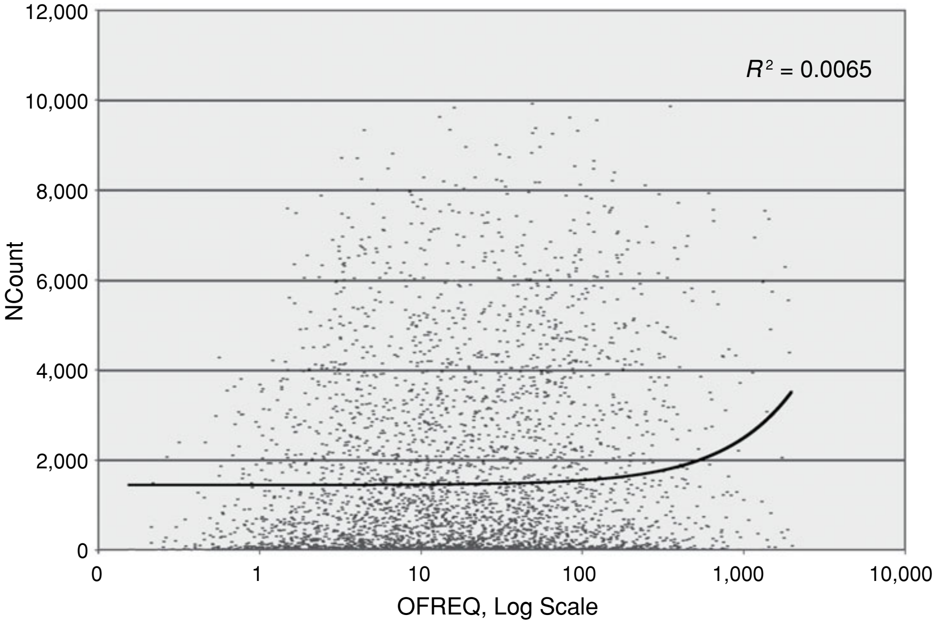

Figure 2. Plot of NCount versus OFREQ for 4,218 words.

the LDRT data collected by Balota et al. (2002) for words and nonwords that average across 100+ subjects' LDRTs per word, we were able to look for a relationship between LDRT and NCount. The work of Buchanan et al. (2001) implies that as NCount decreases, the LDRT should increase, since having many neighbors facilitates lexical access. Using HiDEx, we obtained neighborhoods for the 32,686 words that were common to our lexicon and to the LDRT database at the English Lexicon Project (Balota et al., 2002). The linear correlation between NCount and LDRT was found to be $r = -.29$, $p < .0001$. NUANCE was then used to search for a nonlinear relationship. One was found: LDRT = $\log_{10}$ (NCount + 1) ($r = -.38$, $p < .0001$). This accounted for almost twice the variance of the linear relationship.

Average Radius of Co-occurrence (ARC)

Given a set of neighbors that represents the density of the co-occurrence space around a word, it is possible to calculate the average distance between the words in the neighborhood and the target word. In the cases when the word had no neighbors, the ARC was assigned the distance between the word and its closest neighbor. This measure expresses the proximity of the neighbors to the word in question. As with NCount, we will look for relationships between ARC and OFREQ and ARC and LDRT.

ARC and OFREQ. In a previous section, we found that there was a nonlinear relationship between BSD and OFREQ. We applied the same relationship to ARC, ARC = $1/\sqrt{}$(OFREQ), and found that it correlated at $r = .13$, $r^2 = .02$, $p < .0001$ (see Figure 3). This correlation accounted for 2.5 times less of the variability than did the correlation between BSD and OFREQ, showing that ARC was much less influenced by OFREQ than by BSD.

ARC and LDRT. In the same way, we simulated an experiment for NCount and LDRT, and found a relationship between ARC and LDRT. Using 4,425 words, we found the linear correlation between ARC and LDRT to be $r = .15$, $p < .0001$. This was less than the correlation between NCount and LDRT, and it suggested that further work needs to be done in order to understand the relationship between ARC and LDRT.

Comparison of BSD With ARC and NCount

To test the predictive power of BSD with the influence of frequency removed, we calculated the semipartial correlation between LDRT and BSD that holds 1/SQRT(OFREQ) constant for BSD. This semipartial correlation for the same 4,425 words for which we had HAL BSD measures was $r_{\mathrm{LDRT[BSD,1/SQRT(OFREQ)]}} = .15$, less than the correlation between NCount and LDRT ($-.38$) and the same as the one found for ARC and LDRT.

DISCUSSION

High-dimensional models of context space have been able to reproduce many types of psychological phenomena (Burgess & Lund, 2000). There is still much work to be done in refining and improving the models, as we have attempted to do in this research.

We made two modifications to Lund and Burgess's (1996) HAL model that were then implemented in the HiDEx software. These modifications have allowed us to avoid the problem of being unduly influenced by OFREQ. Furthermore, they have enabled us to find a new measure of semantic density that shows a strong relationship between a statistical model of lexical semantics and a behavioral measure of lexical processing. In the future, we

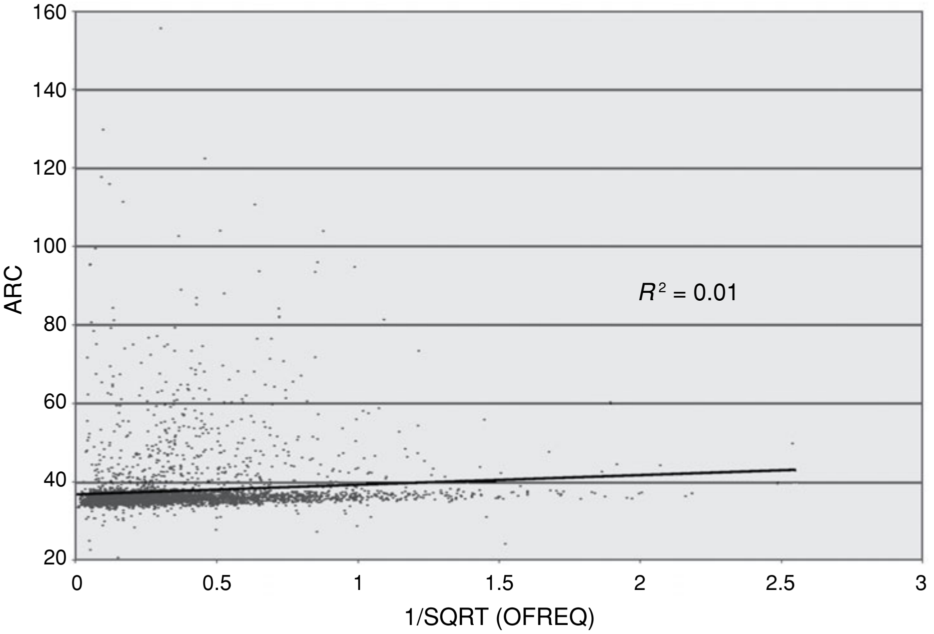

Figure 3. Plot of ARC versus 1/√(OFREQ) or 4,218 words.

hope to vary the model's parameters, and to see how a systematic exploration of the parameter space available in HiDEx will allow us to better understand the relationships among NCount, ARC, and behavioral measures.

REFERENCES

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (CD-ROM). Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Balota, D., Cortese, M., Hutchison, K., Neely, J., Nelson, D., Simpson, G., et al. (2002). *The English Lexicon Project: A Web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords*. St. Louis: Washington University. Retrieved July 7, 2005, from elexicon.wustl.edu.

Buchanan, L., Burgess, C., & Lund, K. (1996). Overcrowding in semantic neighborhoods: Modeling deep dyslexia. *Brain & Cognition*, **32**, 111-114.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, **8**, 531-544.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, **30**, 188-198.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers*, **30**, 272-277.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, **25**, 211-257.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language & Cognitive Processes*, **12**, 177-210.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117-156). Mahwah, NJ: Erlbaum.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval* (pp. 161-169). Las Vegas.

Graff, D. (2002). *The AQUAINT corpus of English news text* (Tech. Rep. No. LDC2002T31). Philadelphia: University of Pennsylvania, Linguistic Data Consortium. (Original work published 1999)

Hollis, G., & Westbury, C. (2006). NUANCE: Naturalistic University of Alberta Nonlinear Correlation Explorer. *Behavior Research Methods*, **38**, 8-23.

Lowe, W. (2001). Toward a theory of semantic space. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 576-581). Mahwah, NJ: Erlbaum.

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

Mohr, G., Stack, M., Ranitovic, I., Avery, D., & Kimpton, M. (2004). Introduction to Heritrix, an archival quality Web crawler. In *4th International Web Archiving Workshop*. Retrieved April 4, 2005, from www.iwaw.net/04/proceedings.php?f=Mohr.

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2004). *An improved method for deriving word meaning from lexical co-occurrence*. Unpublished manuscript. Cambridge, MA: Massachusetts Institute of Technology. Retrieved September 20, 2004, from tedlab.mit.edu/dr/.

Song, D., Bruza, P., & Cole, R. (2004, July 30). *Concept learning and information inferencing on a high-dimensional semantic space*. Paper presented at the ACM SIGIR 2004 Workshop on Mathematical/Formal Methods in Information Retrieval, Sheffield, U.K.

NOTES

[1] Since the OFREQ of words fit best to nonlinear distributions, it is likely that the relationship between the OFREQ of words and other psycholinguistic measures for words are also related in a nonlinear way. Any strong relationship, linear or nonlinear, between OFREQ and a predictor variable indicates that the predictor variable is contaminated, and it is no more than a proxy for OFREQ.

[2] We plan to release HiDEx to the research community in the near future.

[3] The *p* values for all of these relationships are very low, because of the large sample size. Despite this fact, the amounts of variability explained by OFREQ for NCount and BSD are very different.

(Manuscript received November 21, 2005;
revision accepted for publication March 3, 2006.)