SUPPLEMENTARY METHODS

Latent Semantic Analysis

The raw data for LSA begins as a contingency table (often referred to as a term-by-document matrix) in which the columns correspond to all the linguistic episodes, or contexts, that occur in a set of discourse samples (the corpus), and the rows correspond to all the word types that appear in the same samples. The cell values reflect the number of tokens of each word type in each context[1]. Typically, the contexts are semantically coherent passages of text, most often of paragraph size. The result of the analysis is a vector space representation of words and contexts, where the distance between vectors is used as a metric of the semantic similarity between the words and/or contexts represented by the vectors. By way of a very small scale illustration, consider the following text corpus, consisting of a sequence of eight sentences:

S1: "What's your name?" he asked.

S2: "Wendy Moira Angela Darling," she replied with some satisfaction.

S3: "What is your name?"

S4: "Peter Pan."

S5: "Is that all?"

S6: "Yes," he said rather sharply.

S7: She asked where he lived.

S8: "Second to the right," said Peter, "and then straight on till morning."

Taking each sentence as a context, and after removing capitalization and punctuation, the occurrence of

---

[1] It is also possible to perform transformations of the word count per context, called term weighting (see for instance Dumais, 1991).

each of the 36 unique words in the text can be represented in the form of the term-by-document table in Supplementary Methods Table SM1. Looking at the first row we see that the word “what’s” appears only in the first sentence, while the fourth row indicates that the word “he” appears in three different contexts (sentences 1, 6 and 7). In the terminology of matrix algebra, words (rows) are represented by 8-dimensional vectors and contexts (columns) by 36-dimensional vectors.

Supplemental Methods Table SM1. Word-by-context contingency table formed from eight consecutive sentences.

	S1	S2	S3	S4	S5	S6	S7	S8
what's	1	0	0	0	0	0	0	0
your	1	0	1	0	0	0	0	0
name	1	0	1	0	0	0	0	0
he	1	0	0	0	0	1	1	0
asked	1	0	0	0	0	0	1	0
wendy	0	1	0	0	0	0	0	0
moira	0	1	0	0	0	0	0	0
angela	0	1	0	0	0	0	0	0
darling	0	1	0	0	0	0	0	0
she	0	1	0	0	0	0	1	0
replied	0	1	0	0	0	0	0	0
with	0	1	0	0	0	0	0	0
some	0	1	0	0	0	0	0	0
satisfaction	0	1	0	0	0	0	0	0
what	0	0	1	0	0	0	0	0
is	0	0	1	0	1	0	0	0
peter	0	0	0	1	0	0	0	1
pan	0	0	0	1	0	0	0	0
that	0	0	0	0	1	0	0	0
all	0	0	0	0	1	0	0	0
yes	0	0	0	0	0	1	0	0
said	0	0	0	0	0	1	0	1
rather	0	0	0	0	0	1	0	0
sharply	0	0	0	0	0	1	0	0
where	0	0	0	0	0	0	1	0
lived	0	0	0	0	0	0	1	0
second	0	0	0	0	0	0	0	1
to	0	0	0	0	0	0	0	1
the	0	0	0	0	0	0	0	1
right	0	0	0	0	0	0	0	1
and	0	0	0	0	0	0	0	1
then	0	0	0	0	0	0	0	1
straight	0	0	0	0	0	0	0	1
on	0	0	0	0	0	0	0	1
till	0	0	0	0	0	0	0	1
morning	0	0	0	0	0	0	0	1

A straightforward measure of the similarity of two sentences is to just note the number of words in common. Analogously, the degree of similarity between two vectors of equal length can be quantified by calculating their dot product (the sum of the products of their corresponding entries): sentences 1 and 2 contain no words in common, so their dot product similarity is zero. By contrast, sentences 1 and 3 have two words in common ("your" and "name") giving a dot product of 2. An obvious difficulty with this measure is that sentences with more words will be more likely to have overlaps. An alternative similarity measure is to calculate the cosine between the vectors (Salton & Lesk, 1965), which is calculated by normalising the overlap count by the square root of the product of the vector lengths yielding a measure between 0 (no items in common) and 1 (identical). In this example, the cosine between sentence 1 and sentence 2 remains zero, while the cosine between sentence 1 and 3 is 0.45.

Contingency tables formed in this manner are typically very sparse (i.e. very few cells have non-zero values). Moreover, similarities depend on exact matches - so that, for example, the word "animal" would not be considered as a match to the word "mammal". Latent Semantic Analysis (LSA) was proposed by Furnas et al. (1988) as a way of overcoming the limitations of comparing vectors derived from simple contingency tables, by instead measuring similarity in a reduced number of dimensions. Dimension reduction is accomplished by an approach related to principle component analysis known as singular value decomposition (SVD) (Eckart & Young, 1936; Golub & Van Loan, 1996). SVD leads to the calculation of three new matrices, which can be thought of as a 'word space' (the U matrix), a 'context space' (the V matrix), and a diagonal matrix of scaling values (the D matrix). Suppmentary Methods Table SM2 illustrates the matrices from the eight sentence example reduced to 3 dimensions. When appropriately multiplied together these three matrices reproduce an approximation to the original source matrix (Supplementary Methods Table SM3).

Supplementary Methods Table SM2. Word space (U matrix), context space (V matrix) and diagonal matrix of scaling values (D matrix) derived from the first three dimensions obtained using SVD from the term-by-context contingency table in SM1.

U

	1	2	3
what's	-0.01	0.07	-0.23
your	-0.01	0.10	-0.35
name	-0.01	0.10	-0.35
he	-0.06	0.22	-0.50
asked	-0.02	0.18	-0.38
wendy	0.00	0.30	0.14
moira	0.00	0.30	0.14
angela	0.00	0.30	0.14
darling	0.00	0.30	0.14
she	-0.01	0.41	-0.01
replied	0.00	0.30	0.14
with	0.00	0.30	0.14
some	0.00	0.30	0.14
satisfaction	0.00	0.30	0.14
what	0.00	0.03	-0.12
is	0.00	0.03	-0.15
peter	-0.31	-0.02	0.03
pan	-0.03	0.00	0.01
that	0.00	0.00	-0.03
all	0.00	0.00	-0.03
yes	-0.04	0.04	-0.12
said	-0.32	0.02	-0.09
rather	-0.04	0.04	-0.12
sharply	-0.04	0.04	-0.12
where	-0.01	0.11	-0.15
lived	-0.01	0.11	-0.15
second	-0.28	-0.01	0.03
to	-0.28	-0.01	0.03
the	-0.28	-0.01	0.03
right	-0.28	-0.01	0.03
and	-0.28	-0.01	0.03
then	-0.28	-0.01	0.03
straight	-0.28	-0.01	0.03
on	-0.28	-0.01	0.03
till	-0.28	-0.01	0.03
morning	-0.28	-0.01	0.03

D

3.50	0.00	0.00
0.00	3.06	0.00
0.00	0.00	2.81

V

	1	2	3
s1	-0.03	0.22	-0.65
s2	-0.01	0.91	0.39
s3	-0.01	0.08	-0.35
s4	-0.10	-0.01	0.01
s5	0.00	0.01	-0.07
s6	-0.14	0.12	-0.34
s7	-0.03	0.33	-0.43
s8	-0.98	-0.04	0.08

Supplementary Methods Table SM3. Reconstituted source matrix formed by multiplying the three-dimensional matrices (U, D and V shown in Supplementary Methods Table SM2).

	S1	S2	S3	S4	S5	S6	S7	S8
what's	0.47	-0.06	0.24	-0.01	0.05	0.25	0.35	-0.03
your	0.71	-0.12	0.37	-0.01	0.07	0.38	0.52	-0.06
name	0.71	-0.12	0.37	-0.01	0.07	0.38	0.52	-0.06
he	1.06	0.05	0.55	0.00	0.11	0.59	0.83	0.06
asked	0.81	0.08	0.42	-0.01	0.08	0.44	0.64	-0.05
wendy	-0.06	0.97	-0.06	0.00	-0.02	-0.03	0.14	0.00
moira	-0.06	0.97	-0.06	0.00	-0.02	-0.03	0.14	0.00
angela	-0.06	0.97	-0.06	0.00	-0.02	-0.03	0.14	0.00
darling	-0.06	0.97	-0.06	0.00	-0.02	-0.03	0.14	0.00
she	0.29	1.11	0.12	0.00	0.02	0.16	0.43	-0.02
replied	-0.06	0.97	-0.06	0.00	-0.02	-0.03	0.14	0.00
with	-0.06	0.97	-0.06	0.00	-0.02	-0.03	0.14	0.00
some	-0.06	0.97	-0.06	0.00	-0.02	-0.03	0.14	0.00
satisfaction	-0.06	0.97	-0.06	0.00	-0.02	-0.03	0.14	0.00
what	0.24	-0.06	0.13	0.00	0.03	0.13	0.18	-0.02
is	0.29	-0.08	0.15	-0.01	0.03	0.16	0.21	-0.03
peter	-0.04	0.00	-0.03	0.11	-0.01	0.12	-0.02	1.07
pan	-0.01	0.00	0.00	0.01	0.00	0.01	-0.01	0.10
that	0.05	-0.02	0.03	0.00	0.01	0.03	0.03	-0.01
all	0.05	-0.02	0.03	0.00	0.01	0.03	0.03	-0.01
yes	0.25	-0.03	0.13	0.01	0.03	0.15	0.19	0.11
said	0.22	-0.02	0.11	0.10	0.02	0.26	0.17	1.09
rather	0.25	-0.03	0.13	0.01	0.03	0.15	0.19	0.11
sharply	0.25	-0.03	0.13	0.01	0.03	0.15	0.19	0.11
where	0.35	0.14	0.18	-0.01	0.03	0.19	0.29	-0.02
lived	0.35	0.14	0.18	-0.01	0.03	0.19	0.29	-0.02
second	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
to	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
the	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
right	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
and	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
then	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
straight	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
on	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
till	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98
morning	-0.03	0.00	-0.02	0.10	-0.01	0.11	-0.02	0.98

While we chose to define our SVD derived word- and context-spaces using three dimensions for this example, this choice was purely for ease of illustration: larger scale LSA spaces (derived from text corpora consisting of millions of word tokens), are typically defined using hundreds of dimensions; around 300 having been shown to be optimal (Dumais, 1991). The effect of dimension reduction,

however, is the same: semantic relationships are captured among words that do not appear in the same contexts, and among contexts that do not have any words in common, because SVD takes account of indirect co-occurrence (and co-nonoccurrence) data.

How does dimension reduction affect cosine similarity? If we take the cosine between sentence 1 and sentence 2 using columns 1 and 2 from the reconstructed contingency table[2] in Table SM3, the result is −0.028[3], while the cosine between sentences 1 and 3 is 0.995, indicating near identical meaning. Also, the words "what" and "what's" which don't share any contexts now have a strong positive cosine of 0.9958 since the rest of their contexts are so similar in this corpus. Although these results are not surprising in view of the small corpus used, it illustrates how LSA transforms the vector space to represent similarities of meaning. The intuitiveness of the result is characteristic of the output of semantic spaces based on much larger corpora of related texts, which have performed well above chance level when matching synonyms (Landauer & Dumais, 1997) and performing multiple choice tests of semantic knowledge (Landauer et al., 1998)[4].

SVD dimension reduction utilizes all levels of interaction in the original matrix which implies that the most important dimensions are derived not only from the tendency of particular groups of words to co-occur (or not) across contexts, but also from associations between words that appear in *similar* contexts despite never occurring in the *same* context. To illustrate this property, Table SM3 shows the word by

---

[2] Reconstructing an approximation to the original contingency table is done for illustration, and in practice the D and V matrices suffice for context similarity comparisons (see Berry, Dumais, & O'Brien, 1995, or Martin & Berry, 2007 for details).

[3] The cosine can theoretically range from -1.0 to 1.0, but in practice in reduced dimension semantic spaces, the similarity cosine almost always ranges from 0.0 to 1.0 and never dips very deeply into negative values.

[4] Researchers at the University of Colorado at Boulder (USA) have created a number of LSA spaces from digital text collections, including: a 57 million word sample of English and American literature; the texts of three college level psychology textbooks; and groups of 'general reading' texts (novels, newspaper articles, etc.) grouped using their predicted readability by students from third grade (age 8-9) to college level (18+). The spaces based on these text collections are freely available for use with a set of web-based analysis tools for determining similarities among or between terms at http://lsa.colorado.edu (Dennis, 2007).

sentence matrix reconstructed by multiplying the three-dimensional (U, V and D) matrices. Cells with zero entries in the original matrix now contain non-zero (positive and negative values), which indicate the *likelihood* of a word appearing in a particular context given the combination of information about the words and sentences contained in the first three latent variables.

The very technique by which LSA solves the relations between word and passage meanings in order to build its semantic space, namely by using SVD, represents a form of generalization that may have similar mechanisms to cognitive learning and memory formation. Technically, our current understanding of LSA based on SVD does not form a complete model of semantics, since it is not known how to suitably compute SVDs to incrementally add contexts – it is therefore difficult to define a satisfactory learning model, which is a prerequisite for strong attribution of neural underpinnings. Early on in work with LSA (Deerwester et al., 1990), it was recognized that new contexts could be "folded in" to a semantic space without requiring an updated SVD, but these new contexts do not modify the underlying sematic relationships. In the Word Maturity literature where learning is central (e.g. Landauer et al., 2009), this issue is addressed by utilizing multiple semantic spaces each with increasing exposure to contexts to permit computing growth curves, but a more satisfactory, general solution has proven elusive. While LSA can inform understanding semantics, the same underlying matrix decomposition also provides a means of computing the natural modes and frequencies of physical structures such as bridges. While these techniques are powerful and provide useful analogies, careful interpretation is required since we would be unlikely to claim that a bridge rhythmically swinging with the wind is actually computing an SVD.

Supplementary References

Berry, M.W., Dumais, S.T., & O'Brien, G.W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*, 573- 595.

Deerwester, S.C., Dumais, S.T., Landauer, T.K, Furnas, G.W., & Harshman, R.A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science 41* (6), 391-407.

Dennis, S. (2007). How to use the LSA website. In T.K Landauer, D.S. McNamara, S. Dennis, & W. Kintch (Eds.) *Handbook of Latent Semantic Analysis*. pp. 57-70. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Dumais, S. (1991). Improving the retrieval of information from external sources. *Behav Res Meth, Instrum Comput, 23*, 229-236.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1,* 211-218.

Furnas, G.W., Deerwester, S., Dumais, S., Landauer, T.K., Harshman, R.A., Streeter, L.A., et al. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In: Chiaramella Y, editor. SIGIR '88 Proceedings of the11th annual international *ACM SIGIR conference on Research and development in information retrieval*. Grenoble, France: ACM, New York.

Golub, G.H., & Van Loan, C.F. (1996). *Matrix Computations*. Baltimore: Johns Hopkins University Press.

Landauer, T.K, Foltz, P.W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Process, 25*, 259–284.

Landauer, T.K, & Dumais, S.T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychol Rev, 104,* 211–240.

Landauer, T.K, Kireyev, K., & Panaccione, C. (2009). A new yardstick and tool for personalized vocabulary building. In: *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Applications* (27-33). Stroudsburg, PA, Association for Computational Linguistics.

Martin, D.I., & Berry, M.W. Mathematical foundations behind latent semantic analysis. (2007). In T.K Landauer, D.S. McNamara, S. Dennis, & W. Kintch (Eds.) *Handbook of Latent Semantic Analysis*. pp. 35-56. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Salton, G., & Lesk, M. (1965). The SMART Automatic Document Retrieval System - An Illustration. *Comm ACM, 6*, 391-398.