The Journal of Open Source Software

SemanticDistance: Compute Semantic Relatedness between Constituents of Sequential Continuous Text

Jamie Reilly[1,2¶], **Emily B. Myers**[3,4], **Hannah Mechtenberg**[4], and **Jonathan E. Peelle**[5,6,7]

1 Department of Communication Sciences and Disorders, Temple University, United States **2** Department of Psychology and Neuroscience, Temple University, United States **3** Department of Speech, Language, and Hearing Sciences, University of Connecticut, United States **4** Department of Psychological Sciences, University of Connecticut, United States **5** Institute for Cognitive and Brain Health, Northeastern University, United States **6** Department of Communication Sciences and Disorders, Northeastern University, United States **7** Department of Psychology, Northeastern University, United States ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 02 September 2025
Published: unpublished

Summary

SemanticDistance computes pairwise distance between units of language (e.g., word-to-word, ngram-to-word, turn-to-turn) in both structured language samples and unstructured word lists. SemanticDistance has cleaning and formatting options including stopword removal and lemmatization. The package computes two complementary cosine distance indices for each pairwise contrast of interest. SemanticDistance can also be used to examine clustering properties within unstructured word lists, generating dendrograms and simple igraph network plots.

Statment of Need

There are many compelling theoretical and clinical applications for measuring conceptual similarity between words and larger chunks of language (e.g., detection of subtle semantic impairments based on naturalistic language sampling, designing controlled word pair stimuli for language experiments). Although word embeddings such as Word2Vec (Mikolov et al. (2013)), LSA (Landauer et al. (1998)), and GloVE (Pennington et al. (2014)) are widely available, SemanticDistance is unique in its capacity to produce distance metrics in situ within running discourse samples while also offering text cleaning and visualization options.

State of the Field

We know of no software application that computes running measures of semantic distance in continuous unstructured language samples. In addition, SemanticDistance computes different but complementary pairwise semantic distance metrics over the same set of stimuli, allowing researchers to examine semantic flow across continuous stories and narratives.

Software Design

SemanticDistance is an open-source R package designed to accommodate a wide range of user expertise (e.g., neuroscience, linguistics, computer science). The software does not leverage artificial intelligence or use any proprietary dependencies. The package instead indexes an internal lookup database containing publicly available lexical norms.

Research Impact Statement

SemanticDistance has supported one peer-reviewed publication to date in the journal, Cortex (Mechtenberg et al., 2026) along with one refereed conference presentation at the annual conference of the Society for the Neurobiology of Language. Importantly, the application enables a variety of analyses that have been formerly difficult or impossible, including comparing semantic distance measures with other measure of prediction in language, including those generated by large language models.

Description

Semantic distance has been broadly defined an empirical measure of conceptual relatedness between two or more words situated within an n-dimensional space (Reilly et al., 2025). Semantic spaces differ in their dimensionality and biological plausibility. As an R package, SemanticDistance bundles text cleaning and formatting options (e.g., stopword removal, lemmatization) with two two complementary semantic distance metrics that characterize/quantify similarity between pairs of language constituents (e.g., words, ngrams, turns). CosDist_Glo reflects cosine distance between vectors derived from training a GLOVE word embedding model (Pennington et al., 2014). CosDist_SD15 refects cosine distances derived from a 15 dimension sensorimotor and affective space derived from explicit human ratings (Reilly et al., 2023).

SemanticDistance computes indices of relatedness between lexical constituents within the following text formats: 1: **monologues**: stories, narratives, and structured text
2: **dialogues**: turn-to-turn within two-person conversation transcripts.
3: **word pairs in columns**: word pairs
4: **unordered lists**: bags-of-words

One of the most innovative functions of SemanticDistance is its capacity to generate rolling measures of semantic distance within stories or narratives with chunk sizes (e.g., word-to-word, turn-to-turn). Chunk size is an optional argument for many of the software's functions. For example, if a user specified a chunk or n-gram size of 5 words, the software would automatically compute a rollinhg measure of semantic distance betweem each new word in a language sample relative to the five words that preceded it.

AI Usage

SemanticDistance uses an internal database composed of many words with a fixed set of embeddings. The software does not use AI or rely on proprietary package dependencies. We did use GPT 4.0 (OpenAI) to troubleshoot R coding errors and assist with generating complex regular expressions (regex) used for cleaning text data.

References

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Mechtenberg, H., Reilly, J., Peelle, J. E., & Myers, E. B. (2026). Measuring brain sensitivity to semantic distance in spoken narrative comprehension. *Cortex*, *195*, 28–42. https://doi.org/10.1016/j.cortex.2025.12.004

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10/gfshwg

Reilly, J., Finley, A. M., Litovsky, C. P., & Kenett, Y. N. (2023). Bigram semantic distance as an index of continuous semantic flow in natural language: Theory, tools, and applications. *Journal of Experimental Psychology: General*, *152*(9), 2578–2590. https://doi.org/10.1037/xge0001389

Reilly, J., Shain, C., Borghesani, V., Kuhnke, P., Vigliocco, G., Peelle, J. E., Mahon, B. Z., Buxbaum, L. J., Majid, A., Brysbaert, M., Borghi, A. M., De Deyne, S., Dove, G., Papeo, L., Pexman, P. M., Poeppel, D., Lupyan, G., Boggio, P., Hickok, G., … Vinson, D. (2025). What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychonomic Bulletin & Review*, *32*(1), 243–280. https://doi.org/10.3758/s13423-024-02556-7