Exploring Diachronic Lexical Semantics with JESEME

Johannes Hellrich
Graduate School "The Romantic Model.
Variation – Scope – Relevance"
Friedrich-Schiller-Universität Jena
Jena, Germany
johannes.hellrich@uni-jena.de

Udo Hahn
Jena University Language & Information
Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Jena, Germany
udo.hahn@uni-jena.de

Abstract

Recent advances in distributional semantics combined with the availability of large-scale diachronic corpora offer new research avenues for the Digital Humanities. JESEME, the Jena Semantic Explorer, renders assistance to a non-technical audience to investigate diachronic semantic topics. JESEME runs as a website with query options and interactive visualizations of results, as well as a REST API for access to the underlying diachronic data sets.

1 Introduction

Scholars in the humanities frequently deal with texts whose lexical items have become antiquated or have undergone semantic changes. Thus their proper understanding is dependent on translational knowledge from manually compiled dictionaries. To complement this workflow with modern NLP tooling, we developed JESEME,[1] the Jena Semantic Explorer. It supports both lexicologists and scholars with easy-to-use state-of-the-art distributional semantics machinery via an interactive public website and a REST API. JESEME can be queried for change patterns of lexical items over decades and centuries (resources permitting). The website and the underlying NLP pipelines are open source and available via GitHub.[2]

JESEME currently covers five diachronic corpora, two for German and three for English. To the best of our knowledge, it is the first tool ever with such capabilities. Its development owes credits to the interdisciplinary Graduate School "The Romantic Model" at Friedrich-Schiller-Universität Jena (Germany).

2 Related Work

2.1 Distributional Semantics

Distributional semantics can be broadly conceived as a staged approach to capture the semantics of a lexical item in focus via contextual patterns. *Concordances* are probably the most simple scheme to examine contextual semantic effects, but leave semantic inferences entirely to the human observer. A more complex layer is reached with *collocations* which can be identified automatically via statistical word co-occurrence metrics (Manning and Schütze, 1999; Wermter and Hahn, 2006), two of which are incorporated in JESEME as well: Positive pointwise mutual information (PPMI), developed by Bullinaria and Levy (2007) as an improvement over the probability ratio of normal pointwise mutual information (PMI; Church and Hanks (1990)) and Pearson's χ^2 , commonly used for testing the association between categorical variables (e.g., POS tags) and considered to be more robust than PMI when facing sparse information (Manning and Schütze, 1999).

The currently most sophisticated and most influential approach to distributional semantics employs *word embeddings*, i.e., low (usually 300–500) dimensional vector word representations of both semantic and syntactic information. Alternative approaches are e.g., graph-based algorithms (Biemann and Riedl, 2013) or ranking functions from information retrieval (Claveau et al., 2014).

The premier example for word embeddings is skip-gram negative sampling, which is part of the word2vec family of algorithms (Mikolov et al., 2013). The random processes involved in training these embeddings lead to a lack of reliability which is dangerous during interpretation—experiments cannot be repeated without predicting severely different relationships between words (Hellrich and Hahn, 2016a, 2017).

[1] http://jeseme.org
[2] https://github.com/hellrich/JeSemE

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 31–36
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-4006

Word embeddings based on singular value decomposition (SVD; historically popular in the form of Latent Semantic Analysis (Deerwester et al., 1990)) are not affected by this problem. Levy et al. (2015) created $\texttt{SVD}_{\texttt{PPMI}}$ after investigating the implicit operations performed while training neural word embeddings (Levy and Goldberg, 2014). As $\texttt{SVD}_{\texttt{PPMI}}$ performs very similar to `word2vec` on evaluation tasks while avoiding reliability problems we deem it the best currently available word embedding method for applying distributional semantics in the Digital Humanities (Hamilton et al., 2016; Hellrich and Hahn, 2016a).

2.2 Automatic Diachronic Semantics

The use of statistical methods is getting more and more the status of a commonly shared methodology in diachronic linguistics (see e.g., Curzan (2009)). There exist already several tools for performing statistical analysis on user provided corpora, e.g., WORDSMITH[3] or the UCS TOOLKIT,[4] as well as interactive websites for exploring precompiled corpora, e.g., the "advanced" interface for Google Books (Davies, 2014) or DIACOLLO (Jurish, 2015).

Meanwhile, word embeddings and their application to diachronic semantics have become a novel state-of-the-art methodology lacking, however, off-the-shelves analysis tools easy to use for a typically non-technical audience. Most work is centered around `word2vec` (e.g., Kim et al. (2014); Kulkarni et al. (2015); Hellrich and Hahn (2016b)), whereas alternative approaches are rare, e.g., Jo (2016) using `GloVe` (Pennington et al., 2014) and Hamilton et al. (2016) using $\texttt{SVD}_{\texttt{PPMI}}$. Embeddings trained on corpora specific for multiple time spans can be used for two research purposes, namely, screening the semantic evolution of lexical items over time (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016) and exploring the meaning of lexical items during a specific time span by finding their closest neighbors in embedding space. This information can then be exploited for automatic (Buechel et al., 2016) or manual (Jo, 2016) interpretation.

3 Corpora

Sufficiently large corpora are an obvious, yet often hard to acquire resource, especially for diachronic research. We employ five corpora, including the four largest diachronic corpora of acceptable quality for English and German.

The *Google Books Ngram Corpus* (GB; Michel et al. (2011), Lin et al. (2012)) contains about 6% of all books published between 1500 and 2009 in the form of n-grams (up to pentagrams). GB is multilingual; its English subcorpus is further divided into regional segments (British, US) and genres (general language and fiction texts). It can be argued to be not so useful for Digital Humanities research due to digitalization artifacts and its opaque and unbalanced nature, yet the English Fiction part is least effected by these problems (Pechenick et al., 2015; Koplenig, 2017). We use its German (GB German) and English Fiction (GB fiction) subcorpora.

The *Corpus of Historical American English*[5] (COHA; Davies (2012)) covers texts from 1800 to 2009 from multiple genres balanced for each decade, and contains annotations for lemmata.

The *Deutsches Textarchiv*[6] (DTA, 'German Text Archive'; Geyken (2013); Jurish (2013)) is a German diachronic corpus and consists of manually transcribed books selected for their representativeness and balance between genres. A major benefit of DTA are its annotation layers which offer both orthographic normalization (mapping archaic forms to contemporary ones) and lemmatization via the CAB tool (Jurish, 2013).

Finally, the *Royal Society Corpus* (RSC) contains the first two centuries of the *Philosophical Transactions of the Royal Society of London* (Kermes et al., 2016), thus forming the most specialized corpus in our collection. Orthographic normalization as well as lemmatization information are provided, just as in DTA. RSC is far smaller than the other corpora, yet was included due to its relevance for research projects in our graduate school.

4 Semantic Processing

The five corpora described in Section 3 were divided into multiple non-overlapping temporal slices, covering 10 years each for COHA and the two GB subcorpora, 30 years each for the smaller DTA and finally two 50 year slices and one 19 year slice for the even smaller RSC (as

[3] `http://lexically.net/wordsmith`

[4] `http://www.collocations.de/software.html`

[5] `http://corpus.byu.edu/coha/`

[6] TCF version from May 11th 2016, available via `www.deutschestextarchiv.de/download`

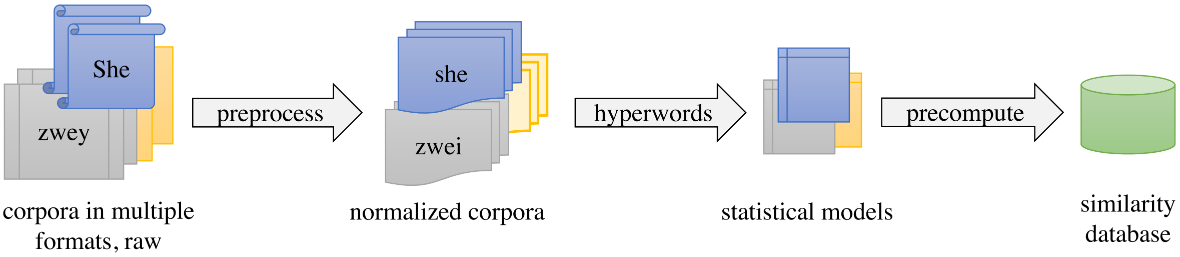


Figure 1: Diagram of JESEME's processing pipeline.

provided in the corpus, roughly similar in size). We removed non-alphanumeric characters during pre-processing and transformed all English text to lowercase. Lemmata were used for the stronger inflected German (provided in DTA, respectively a mapping table created with the CAB webservice (Jurish, 2013) for the German GB subcorpus) and the rather antiquated RSC (provided in the corpus).

We calculated `PPMI` and χ^2 for each slice, with a context window of 4 words, no random sampling, context distribution smoothing of 0.75 for `PPMI`, and corpus dependent minimum word frequency thresholds of 50 (COHA, DTA and RSC) respectively 100 (GB subcorpora).[7] The `PPMI` matrices were then used to create $\texttt{SVD}_{\texttt{PPMI}}$ embeddings with 500 dimensions. These calculations were performed with a modified version of HYPERWORDS[8] (Levy et al., 2015), using custom extensions for faster pre-processing and χ^2. The resulting models have a size of 32 GB and are available for download on JESEME's `Help` page.[9]

To ensure JESEME's responsiveness, we finally pre-computed similarity (by cosine between word embeddings), as well as context specificity based on `PPMI` and χ^2 . These values are stored in a POSTGRESQL[10] database, occupying about 60GB of space. Due to both space constraints (scaling with $\mathcal{O}(n^2)$ for vocabulary size n) and the lower quality of representations for infrequent words, we limited this step to words which were among the 10k most frequent words for all slices of a corpus, resulting in 3,1k – 6,5k words per corpus. In accordance with this limit, we also discarded slices with less than 10k (5k for RSC) words above the minimum frequency threshold used during `PPMI` and χ^2 calculation, e.g., the 1810s and 1820s COHA slices. Figure 1 illustrates this sequence of processing steps, while Table 1 summarizes the resulting models for each corpus.

Corpus	Years	Words
COHA	1830–2009	5,101
DTA	1751–1900	5,338
GB Fiction	1820–2009	6,492
GB German	1830–2009	4,449
RSC	1750-1869	3,080

Table 1: Years and number of words modelled for each corpus in JESEME.

5 Website and API

JESEME provides both an interactive website and an API for querying the underlying database. Both are implemented with the SPARK[11] framework running inside a JETTY[12] Web server.

On JESEME's initial landing page, users can enter a word into a search field and select a corpus. They are then redirected to the result page, as depicted in Figure 2. Query words are automatically lowercased or lemmatized, depending on the respective corpus (see Section 4). The result page provides three kinds of graphs, i.e., `Similar Words`, `Typical Context` and `Relative Frequency`.

`Similar Words` depicts the words with the highest similarity relative to the query term for the first and last time slice and how their similarity values changed over time. We follow Kim et al. (2014) in choosing such a visualization, while we refrain from using the two-dimensional projection used in other studies (Kulkarni et al., 2015; Hamilton et al., 2016). We stipulate that the latter could

[7] Parameters were chosen in accordance with Levy et al. (2015) and Hamilton et al. (2016).

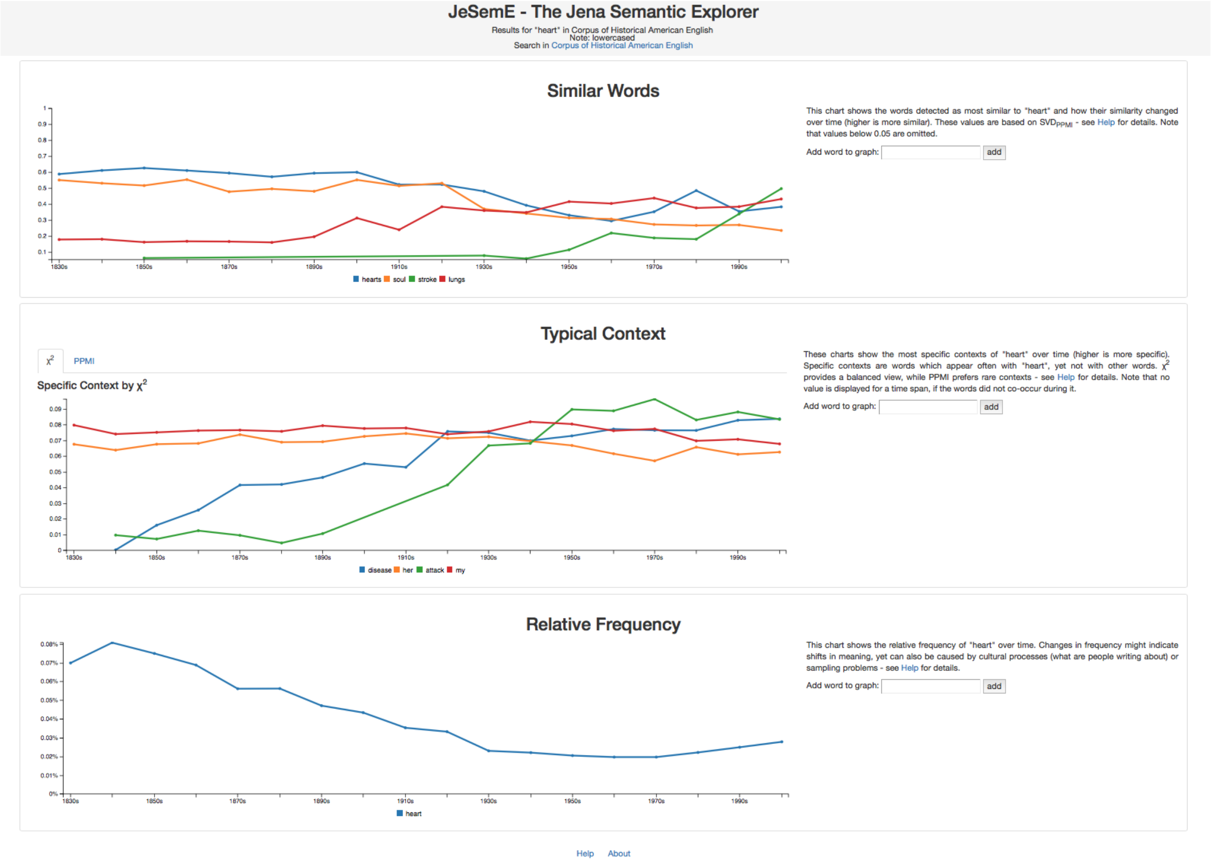
[8] `https://bitbucket.org/omerlevy/hyperwords`

[9] `http://jeseme.org/help.html#download`

[10] `https://www.postgresql.org`

[11] `http://sparkjava.com`

[12] `http://www.eclipse.org/jetty`



JeSemE - The Jena Semantic Explorer

Results for "heart" in Corpus of Historical American English
Note: lowercased
Search in Corpus of Historical American English

Similar Words

This chart shows the words detected as most similar to "heart" and how their similarity changed over time (higher is more similar). These values are based on SVD$_{PPMI}$ - see Help for details. Note that values below 0.05 are omitted.

Add word to graph: add

Typical Context

Specific Context by χ^2

These charts show the most specific contexts of "heart" over time (higher is more specific). Specific contexts are words which appear often with "heart", yet not with other words. χ^2 provides a balanced view, while PPMI prefers rare contexts - see Help for details. Note that no value is displayed for a time span, if the words did not co-occur during it.

Add word to graph: add

Relative Frequency

This chart shows the relative frequency of "heart" over time. Changes in frequency might indicate shifts in meaning, yet can also be caused by cultural processes (what are people writing about) or sampling problems - see Help for details.

Add word to graph: add

Help About

Figure 2: Screenshot of JESEME's result page when searching for the lexical item *"heart"* in COHA.

be potentially misleading by implying a constant meaning of those words used as the background (which are actually positioned by their meaning at a single point in time).

`Typical Context` offers two graphs, one for χ^2 and one for `PPMI`, arranged in tabs. Values in typical context graphs are normalized to make them comparable across different metrics.

Finally, `Relative Frequency` plots the relative frequency measure against all words above the minimum frequency threshold (see Section 4). All graphs are accompanied by a short explanation and a form for adding further words to the graph under scrutiny. The result page also provides a link to the corresponding corpus, to help users trace JESEME's computational results.

As an example, consider JESEME's search for *"heart"* in COHA as depicted in Figure 2. The `Similar Words` graph depicts a lowered similarity to *"soul"* and increased similarity to *"lungs"*, and more recently also *"stroke"*, which we interpret as a gradual decrease in metaphorical usage. Since COHA is balanced, we assume this pattern to indicate a true semantic change; a similar change is also observable in the GB Fiction dataset, yet not in the highly domain-specific RSC. Note that this change is unlikely to be linked with the decreased frequency of *"soul"*, as `PMI`-derived metrics are known to be biased towards infrequent words (Levy et al., 2015). This shift in meaning is also visible in the `Typical Context` graphs, with *"attack"* and *"disease"* being increasingly specific by both χ^2 and `PPMI`. Note that metaphorical or metonymical usage of *"heart"* is historically quite common (Niemeier, 2003), despite its long-known anatomical function (Aird, 2011).

The database underlying JESEME's graphs can also be queried via a `REST` API which provides JSON encoded results. API calls need to specify the corpus to be searched and one (frequency) or two (similarity, context) words as `GET` parameters.[13] Calling conventions are further detailed on JESEME's `Help` page.[14]

[13] For example `http://jeseme.org/api/similarity?word1=Tag&word2=Nacht&corpus=dta`

[14] `http://jeseme.org/help.html#api`

6 Conclusion

We presented JESEME, the Jena Semantic Explorer, an interactive website and REST API for exploring changes in lexical semantics over long periods of time. In contrast to other corpus exploration tools, JESEME is based on cutting-edge word embedding technology (Levy et al., 2015; Hamilton et al., 2016; Hellrich and Hahn, 2016a, 2017) and provides access to five popular corpora for the English and German language. JESEME is also the first tool of its kind and under continuous development.

Future technical work will add functionality to compare words **across** corpora which might require a mapping between embeddings (Kulkarni et al., 2015; Hamilton et al., 2016) and provide optional stemming routines. Both goals come with an increase in precomputed similarity values and will thus necessitate storage optimizations to ensure long-term availability. Finally, we will conduct a user study to investigate JESEME's potential for the Digital Humanities community.

Acknowledgments

This research was conducted within the Graduate School *"The Romantic Model. Variation – Scope – Relevance"* supported by grant GRK 2041/1 from *Deutsche Forschungsgemeinschaft (DFG)* (http://www.modellromantik.uni-jena.de/).

References

William C. Aird. 2011. Discovery of the cardiovascular system: from Galen to William Harvey. *Journal of Thrombosis and Haemostasis* 9(s1):118–129.

Chris Biemann and Martin Riedl. 2013. Text: now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling* 1(1):55–95.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016. Feelings from the past: adapting affective lexicons for historical emotion analysis. In *LT4DH — Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities @ COLING 2016. December 11, 2016, Osaka, Japan.* pages 54–61.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods* 39(3):510–526.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Vincent Claveau, Ewa Kijak, and Olivier Ferret. 2014. Improving distributional thesauri by exploring the graph of neighbors. In *COLING 2014 – Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland, August 23-29, 2014*. pages 709–720.

Anne Curzan. 2009. Historical corpus linguistics and evidence of language change. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin; New York/NY, volume 2 of *Handbooks of Linguistics and Communication Science, 29*, pages 1091–1109.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora* 7(2):121–157.

Mark Davies. 2014. Making Google Books n-grams useful for a wide range of research on language change. *International Journal of Corpus Linguistics* 19(3):401–416.

Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Alexander Geyken. 2013. Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv. In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, Berlin-Brandenburgische Akademie der Wissenschaften, number 4 in Thesaurus Linguae Aegyptiae, pages 221–234.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Long Papers. Berlin, Germany, August 7-12, 2016*. pages 1489–1501.

Johannes Hellrich and Udo Hahn. 2016a. Bad company: Neighborhoods in neural embedding spaces considered harmful. In *COLING 2016 — Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, December 11-16, 2016*. pages 2785–2796.

Johannes Hellrich and Udo Hahn. 2016b. Measuring the dynamics of lexico-semantic change since the German Romantic period. In *Digital Humanities 2016 — Conference Abstracts of the 2016 Conference of the Alliance of Digital Humanities Organizations (ADHO). 'Digital Identities: The Past and the Future'. Kraków, Poland, 11-16 July 2016*. pages 545–547.

Johannes Hellrich and Udo Hahn. 2017. Don't get fooled by word embeddings: better watch their neighborhood. In *Digital Humanities 2017 — Conference Abstracts of the 2017 Conference of the Alliance of Digital Humanities Organizations (ADHO). Montréal, Quebec, Canada, August 8-11, 2017*.

Eun Seo Jo. 2016. Diplomatic history by data. Understanding Cold War foreign policy ideology using networks and NLP. In *Digital Humanities 2016 — Conference Abstracts of the 2016 Conference of the Alliance of Digital Humanities Organizations (ADHO). 'Digital Identities: The Past and the Future'. Kraków, Poland, 11-16 July 2016*. pages 582–585.

Bryan Jurish. 2013. Canonicalizing the Deutsches Textarchiv. In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*. Berlin-Brandenburgische Akademie der Wissenschaften, number 4 in Thesaurus Linguae Aegyptiae, pages 235–244.

Bryan Jurish. 2015. DiaCollo: on the trail of diachronic collocations. In *Proceedings of the CLARIN Annual Conference 2015. Book of Abstracts. Wroław,, Poland, 14-16 October, 2015*. pages 28–31.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: from uncharted data to corpus. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*. pages 1928–1931.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the Workshop on Language Technologies and Computational Social Science @ ACL 2014. Baltimore, Maryland, USA, June 26, 2014*. pages 61–65.

Alexander Koplenig. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets: reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities* 32(1):169–188.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *WWW '15 — Proceedings of the 24th International Conference on World Wide Web: Technical Papers. Florence, Italy, May 18-22, 2015*. pages 625–635.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27 — Proceedings of the Annual Conference on Neural Information Processing Systems 2014. Montréal, Quebec, Canada, December 8-13, 2014*. pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3:211–225.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram Corpus. In *ACL 2012 — Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Jeju Island, Korea, July 10, 2012*. pages 169–174.

Chris Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, chapter 5: Collocations.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR 2013 — Workshop Proceedings of the International Conference on Learning Representations. Scottsdale, Arizona, USA, May 2-4, 2013*.

Susanne Niemeier. 2003. Straight from the heart: metonymic and metaphorical explorations. In Antonio Barcelona, editor, *Metaphor and Metonymy at the Crossroads: A Cognitive Perspective*, Mouton de Gruyter, Berlin; New York/NY, number 30 in Topics in English Linguistics, pages 195–211.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books Corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One* 10(10):e0137041.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: global vectors for word representation. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, October 25-29, 2014*. pages 1532–1543.

Joachim Wermter and Udo Hahn. 2006. You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction. In *COLING-ACL 2006 — Proceedings of the 21st International Conference on Computational Linguistics & 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, 17-21 July 2006*. volume 2, pages 785–792.