# How a Stylistic, Machine-Generated Thesaurus Impacts a Writer's Process

**Katy Ilonka Gero**
Columbia University
New York City, NY 10027, USA
katy@cs.columbia.edu

**Lydia B. Chilton**
Columbia University
New York City, NY 10027, USA
chilton@cs.columbia.edu


*C&C '19*, June 23–26, 2019, San Diego, CA, USA
ACM 978-1-4503-5917-7/19/06.
https://doi.org/10.1145/3325480.3326573

## Abstract

Writers regularly use a thesaurus to help them write well; the thesaurus is one of the few widespread writing support tools and many writers find it integral to their writing practice. A normal thesaurus is hand-crafted and structured around strict synonymy for a given word sense. However, writers rarely look for a perfectly synonymous word – instead they have additional ideas or constraints, such as words that are less cliche, more specific, or less gendered. Poets describe their usage as searching for words that "hold more interesting connotations." We present a machine learning approach to thesaurus generation, using word embeddings, that leverages stylistically distinct corpora – such as naturalist writing, novels by a particular author, or writing from a technical discipline. We show examples of how stylistic thesauruses differ from each other and from a regular thesaurus, as well as preliminary responses from two writers who are given multiple stylistic thesauruses. Writers describe these thesauruses as reflective of style, unique from each other, and more exploratory and associative than a regular thesaurus. They also describe an increased attention to connotation. We outline plans for quantitative evaluation of stylistic thesauruses, and user studies to understand their impact on specific tasks.

## CCS Concepts
**•Human-centered computing → Interactive systems and tools; •Applied computing → Arts and humanities;**

## Author Keywords
Creativity support tools; writing support; natural language processing; human-computer interaction.

## Introduction
The thesaurus has become an integral writing support tool, distinct from dictionary definitions, used by journalists, scientists, and poets alike across a wide range of writing tasks and genres. Automatic synonym discovery has a long history in information retrieval, where synonyms allow search queries to be semantically expanded [12, 11]. But little work has used a computational approach to improve the thesaurus as a writing tool.

In preliminary interviews with writers, we found that writers rarely express preferences for the different thesauruses available to them. Instead they opt for whatever they consider the 'default', whether this be the thesaurus built into their operating system, or the results from querying a search engine. Many writers consider the thesaurus to be an important part of their writing process.

Yet most talked at length about the difficulty they have in finding words, and the variety of constraints or goals they have for these words. Strict synonymy, or even a drop-in replacement word, is not essential for most writers, who instead may use a thesaurus to expand an idea or discover a new one. For instance, a journalist may be looking for a less gendered word, a scientist for a more specific word, or a poet for a less cliche word. In each case, the desired word actually reflects a new meaning or connotation the writer hopes to express.

Recently, word embeddings have become popular in the natural language processing community due to their accurate representation of word-level semantics [8, 9]. Word embeddings are vector representations of words, where each vector represents a word being "embedded" in a high dimensional space (normally around 200 dimensions). Words that are near each other in this space are semantically similar. For example, "computer" and "program" would be near each other, whereas "computer" and "cat" would not. These word embeddings can be learned from a corpus of text using a neural network architecture [7, 9, 4].

Other work on writing support has also had to contend with style. Systems that make suggestions by finishing a writer's sentence do better when focused on a particular style [6]. Tools that scaffold a writing task often focus on narrow domains, like writing an email to request help [2] or writing a newspaper article [5]. InkWell [1] assists writers by showing them stylistic variations of a text.

In this work, we leverage word embeddings and part-of-speech tagging to create domain-specific thesauruses that reflect stylistic differences between fields, genres, and even individual writers. In exploratory studies, we interrogate the following research questions:

- How can we generate stylistic thesauruses?
- How are machine-generated thesauruses, which use distinct corpuses of text, different from each other? How are they different from a general purpose thesaurus?
- How does having one or more custom thesauruses, which prioritize style over meaning, impact a writer's process?

## System design

Our goal is to generate a custom thesaurus from a corpus of text, where the words returned by this custom thesaurus reflect the style presented in the corpus.

To retrieve stylistically distinct words, we first select a stylistically distinct textual corpus to learn from. In order to understand how well the approach would work across different domains, we pick one technical domain (science paper abstracts), one literary domain (James Joyce novels, known for their distinctive stream-of-consciousness style writing), and one domain somewhere in the middle (Charles Darwin's naturalist writing, which is scientific in nature but was published for a general audience). Here are three the corpuses we use as test cases, though our algorithm is extendable to any corpus:

- 40k math and science abstracts from arXiv.org
- the collected works of James Joyce[1]
- the collected works of Charles Darwin

For each corpus, we learn a set of word embeddings for every word that occurs more than 5 times. For this paper, all our examples and evaluations come from using the word2vec algorithm [7], with a context window of 5. We use the Gensim[2] Python library implementation.

However, we have also experimented with different algorithms. In one case, instead of using a linear context window, in which the context words used in training are the words on either side of the target word, we use the words closest in the dependency parse tree [3]. In another case, we use structured embeddings, that train embeddings for each corpus in a joint high-dimensional space, allowing for easier comparison [10]. In the future we plan to evaluate the differences between these algorithms with a combination of human evaluators and automatic evaluation.

From a word embedding, we are able to retrieve the top $n$ words closest to any word in the vocabulary. To improve the quality and relevance of the returned words, we also filter the top $n$ words by part-of-speech. To do so, we first parse each corpus, tagging every word with its part-of-speech using the Spacy parser[3]. Then, we build a dictionary for each corpus of every word in the vocabulary, and all the parts-of-speech it was used as in that corpus. When searching for a word, if a user does not specify a part-of-speech, one is randomly selected.

## Differences between thesauruses

We present an exploratory study of the differences between our machine-generated thesauruses as well as a hand-crafted, general purpose thesaurus. We look at query results from thesauruses generated from the corpuses described in the System Design section (Science, Joyce, Darwin) plus the results from the most commonly used general purpose online thesaurus, based on our preliminary interviews with writers, Thesaurus.com. Thesaurus.com uses Roget's 21st Century Thesaurus, Third Edition Copyright.

We examine the results for two common words: the word 'work' is an overused word in professional writing[4] and the word 'look' is an overused word in poetry writing classes[5].

[1] We retrieved both the Joyce and Darwin corpuses from Project Gutenberg using Allison Parrish's Python library. https://github.com/aparrish/gutenberg-dammit

[2] https://radimrehurek.com/gensim/

[3] https://spacy.io/

[4] "20 Overused Words Grammarly Can Help You Diversify in Your Writing" https://www.grammarly.com/blog/common-synonyms/

[5] Discussion with poetry workshop instructor.

query: work (noun)

- **Science** paper, article, chapter, thesis, manuscript, research, survey, dissertation
- **Joyce** chance, marry, care, trust, talk, drink, somebody, bear
- **Darwin** subject, conclusions, views, observation, hypothesis, practice, observations, discussion
- **Thesaurus.com** effort, endeavor, industry, job, performance, production, struggle, task

query: look (verb)

- **Science** arrive, aim, begin, succeed, go, move, fall, thin
- **Joyce** keep, live, come, speak, use, leave, eat, meet
- **Darwin** marvel, turn, consider, return, feel, extend, treat, take
- **Thesaurus.com** consider, glance, notice, peer, read, see, stare, study

We will only discuss the results for "work".

The **Science** thesaurus results mostly in words associated with how science presents work: papers, articles, theses, etc. This is a very particular conception of work that accurately reflects the primary meaning of work in the sciences. It's interesting to note that it doesn't include words like "experiments" or "writing". Perhaps in abstracts (the underlying corpus) "work" is primarily used when referring to the work of others, which is instantiated as reports on work, rather than the actual undertaking itself.

Compare this with the **Darwin** thesaurus, where work refers to the kind of work Darwin did, such as observations, views, and conclusions. Here we see Darwin's focus on the content of his naturalist work instead of the write-up. Notably this thesaurus returns "observation" but not "experiment".

**Joyce** is very different from either of these, with words that don't have clear synonymous relations to "work", such as "chance", and "marry". These results represent of a free association with work that is characteristic of Joyce's writing, making the writer think of how "work" relates to "chance" – perhaps how our work is impacted or ultimately guided by chance – and how "work" relates to "marriage".

All of these are distinct from from **Thesaurus.com**, which focus on "work" as "labor". Because it is general purpose, it does not bring to the surface any connotations of what work might mean to a particular person, genre, or field.

## Impact on writers

*Interviews about regular thesaurus use*

The first author conducted semi-structured interviews with four writers currently or recently enrolled in an MFA program with a concentration in poetry. Writers were asked if they used dictionaries or thesauruses and in what way. The were also asked to give an example from the past week of a specific usage and the surrounding context and results.

Two of the four writers primarily discussed using both a dictionary and thesaurus to look for more correct or suitable words. A dictionary could confirm correct usage; conversely a thesaurus could suggest a more exact or more common word. (Sometimes writers did want to use more esoteric words; at other times they opted for words most readers would be familiar with.) One of these writers regularly used reverse look up dictionaries to find technical words, as well as rhyming dictionaries to find words with specific phonetic properties.

The other two writers specifically mentioned using a thesaurus to look for more "interesting" words. One of these writers used the thesaurus during the writing process – as opposed to during editing – to find words that "hold more interesting connotations". Because they would do this while writing a first draft, a thesaurus sometimes impacted the direction of their work by presenting an unexpected word. The other writer discussed at length their process for replacing nouns or nouns phrases they thought were too cliche, by thinking of the connotations of those words and thinking of other words that hold those connotations, or replacing verbs, by thinking of verbs that were unexpected or contrastive given the subject or direct object of the verb.

From this we found that writers rarely use a thesaurus to find a strict synonym, but rather have a diversity of other goals that supersede synonymy, such as specificity, generality, cliche-ness, phonetics, connotation, and contrast with other parts of the phrase or sentence.

These additional constraints are often considered collectively as 'writing style', in which particular authors, genres, or domains have a shared desire to meet certain constraints. For instance, most poets share a disdain for the cliche, while scientists are not worried that a technical term has been overused. In contrast, most scientists aspire for highly specific technical words, while poets are more willing to rely on connotation. Our work addresses these user needs at this level of 'style', instead of the lower level of individual, specific constraints.

*Preliminary responses to custom thesauruses*
To gauge the effectiveness of our custom thesauruses, we showed them to two writers who have a range of experience writing essays and news articles, as well as fiction and poetry. They were shown the thesauruses embedded in a simple web application such that they could easily query them, as well as some example queries. They were then asked to play around with the application, thinking aloud about what they thought the results reflected, and how it might be integrated into their writing practice.

The first writer noted that they would reflect on why a word was returned. For instance, when querying the word "flash", they wondered about how Joyce was using the word such that "violet" turned up. Perhaps "flash" was often used as "a flash of color", implying that "violet" was a common color in Joyce's work. They thought the results indicated different connotations, and what is related to those connotations. For instance, when querying the word "work", the Science corpus returns words like "paper", "chapter", and "manuscript", indicating that in science writing "work" mostly connotes an object or result, rather than an activity. They thought it would be particularly useful for writing in a certain voice or character, or for coming up with thematically exciting words. They wondered what kind of thesaurus would come from a corpus of nautical novels (like Moby Dick).

The second writer started to associate each corpus with a set of adjectives to describe its style. They described the Joyce results as "flowery", "pretentious", and "intellectual". They were drawn to the Joyce results as they were often more unexpected. For example, when querying "stylish" they found "opulent" and "muffled". Although these words were not particularly synonymous, they created an interesting association and spurred new ideas. They were particularly interested in using this to discover how word usage differed across cultures. The second writer is from New Zealand, and reflected that upon moving to the U.S. there was a language barrier, especially in writing essays that drew on more particular notions of fluid writing.

## Conclusion and future work

We present a novel approach to creative writing support through automatic thesaurus creation. We use word embeddings and part-of-speech tagging to turn any textual corpus into a unique thesaurus, and show examples of how different corpora return different and relevant results. As a preliminary study, we created a web application that allowed writers to query several custom thesauruses at once, and had two writers use and reflect on this tool. Both writers found the results serendipitous and inspiring.

In the future, we plan to compare different algorithms using human and automatic evaluations to understand how well they result in synonymous, related, and corpus-specific words. Automatic evaluations could calculate how specific returned words are to a given corpus (compared to other corpora), how similar the top $n$ words are to the query word, and how large a corpus must be to produce consistent results. Human evaluators could indicate how synonymous and related returned words are to the query word, if returned words are more specific or general, and if returned words are specific to the given corpus.

We also plan to run user studies in which writers use one or several stylistic thesauruses to do a specific writing task, such as re-writing a scientific abstract as a short blog post, or re-writing a poem to be based on a different central image. Using a think-aloud method [13] we could quantify the interaction dynamics and better understand the way stylistic thesauruses impact cognition.

## Acknowledgements

Katy Ilonka Gero is supported by an NSF GRF (DGE - 1644869).

## REFERENCES


[1] Richard P Gabriel, Jilin Chen, and Jeffrey Nichols. 2015. InkWell: A Creative Writer's Creative Assistant. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. ACM, 93–102.

[2] Julie S. Hui, Darren Gergle, and Elizabeth M. Gerber. 2018. IntroAssist: A Tool to Support Writing Introductory Help Requests. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 22, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3173596

[3] Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 302–308.

[4] 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 (2015), 211–225. DOI: http://dx.doi.org/10.1162/tacl_a_00134

[5] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the News: Digital Creativity Support for Journalists. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 475, 11 pages. DOI: http://dx.doi.org/10.1145/3173574.3174049

[6] Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. 2017. Synthetic literature: Writing science fiction in a co-creative process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. 29–37.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[8] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 746–751.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[10] Maja Rudolph, Francisco Ruiz, Susan Athey, and David Blei. 2017. Structured embedding models for grouped data. In *Advances in Neural Information Processing Systems*. 251–261.

[11] Hinrich Schütze and Jan O Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management* 33, 3 (1997), 307–318.

[12] Peter D. Turney. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Machine Learning: ECML 2001*, Luc De Raedt and Peter Flach (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 491–502.

[13] MW Van Someren, YF Barnard, and JAC Sandberg. 1994. *The think aloud method: a practical approach to modelling cognitive*. Citeseer, Chapter 3.