Investigating Homographic Pun Distances via Semantic Embeddings and NLP

Srikanth Iyer, Annajirao Challa
siyer5@binghamton.edu, achalla@binghamton.edu

***Abstract* - This project delves into the complex nature of puns through the lens of Semantic Diversity (SemD). SemD quantifies the ambiguity of words by evaluating their variability in contextual usage, offering a departure from conventional methods reliant on dictionary definitions. We initiated an investigation by calculating SemD values for the literary compositions of renowned authors, including William Shakespeare, Gilbert Keith Chesterton, Oscar Wilde, and others. Our analysis aims to unravel the variability of word meanings within the body of work of individual authors. A higher SemD value signifies the presence of a word in a myriad of diverse contexts, suggesting heightened ambiguity. By scrutinizing SemD in relation to puns, we shed light on the intricate interplay between linguistic ambiguity and comedic wordplay. Our findings aimed to look at the semantic diversity of the words used in the compositions and their underlying nuances.**

Keywords: Puns, Linguistics, Literature, Semantic Diversity

INTRODUCTION

Words in natural languages are often polysemous, meaning they have multiple related but distinct meanings depending on the context. This one-to-many mapping of form to meaning presents a significant challenge for NLP tasks like machine translation, human-computer interaction, and text understanding.

Traditional word sense disambiguation (WSD) approaches assume there is a single intended meaning for every word, failing to account for cases where ambiguity is a function of the word and cases where this ambiguity was deployed deliberately. In literature, polysemy is a powerful stylistic device that allows authors to imbue words with multiple layers of meaning, creating ambiguity, suspense, and depth. The intentional use of lexical ambiguity through punning is a particularly common source of humor. Puns exploit different meanings of a term, either within the same lexical category (homonymy) or across categories (polysemy), to set up a conflict between the expected and actual interpretations. Resolving this conflict by recognizing the alternative meaning is key to understanding humor.

Measuring the semantic distance between the polysemous interpretations of a pun can provide insights into the cognitive processes involved in comprehending and appreciating the humor. It allows quantifying the conceptual leap or incongruity that the reader must overcome. This semantic distance impacts the perceived funniness, processing effort, and overall reader experience with the text. By focusing on homographic puns that use words spelled identically but with different pronunciations and meanings (e.g. "the buck stops here" referring to both a male deer and money), the project aims to develop computational methods to map the semantic spaces of different literary works and evaluate how semantic distances contribute to the nature of comedy versus non-comedic writing [6].

LINGUISTIC KEYWORDS ON PUNS

Homography refers to words that are spelled the same but have different meanings, regardless of whether they sound alike. This is distinct from homophony, where words sound the same but may be spelled differently and have different meanings [1]. A **homophone** is a word that shares the same pronunciation as another word but has a different meaning [2]. These words might have identical spellings, like "bat" (flying mammal) and "bat" (sports equipment), or they may be spelled differently, such as "write," "right," and "rite." Additionally, the term "homophone" can extend to longer or shorter linguistic units, such as phrases, individual letters, or groups of letters, which sound the same as their counterparts. Homophones that are spelled the same are deemed both homographs and homonyms. For example, the word "read," as in "He is well read" (meaning he is very learned), versus the sentence "I read that book" (meaning I have finished reading that book). Homophones that are spelled differently are also called **heterographs**. For instance, "to," "too," and "two".

In the field of linguistics, **homonyms** encompass words that fall into two categories: homographs, which are words sharing identical spellings regardless of pronunciation, and homophones, which share the same pronunciation irrespective of spelling. Some words can belong to both categories.

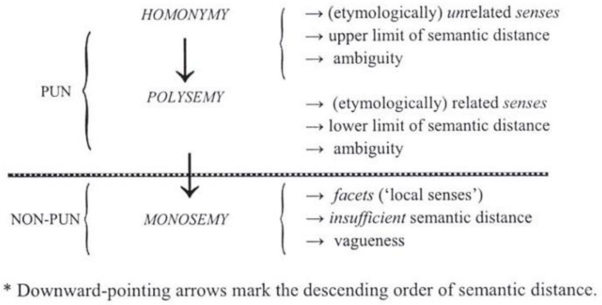
Paronyms, on the other hand, are words that sound or are written similarly but possess distinct lexical meanings [3]. They contrast with homonyms, which are words with different meanings that have the same pronunciation or spelling. Examples of English paronyms include Affect and Effect, Alternately and Alternatively, Collision and Collusion, Complement and Compliment, Conjuncture and Conjecture, Continuous and Contiguous, Deprecate and Depreciate, Eclipse and Ellipse, Elicit and Illicit, Excise and Exercise, Principle and Principal, Stationary and Stationery, Upmost and Utmost.

A **heteronym** refers to a word that shares the same spelling with another word but has a distinct pronunciation and meaning [4]. These words are homographs but not homophones. For instance, "tear" (as in crying) and "tear" (a rip) are heteronyms, but "wind" (a gust of air) and "wind" (to twist) are not, as they share the same pronunciation. Heteronyms can vary in their pronunciation due to differences in vowel sounds, stress patterns, or other linguistic factors. Homographs and Heteronyms are common forms of Puns in literature.

LITERATURE REVIEW

In 1974, Leech introduced a semantic model with multiple layers, distinguishing between conceptual meaning, referred to as "sense," and other types of meaning categorized under "communicative value" [4]. This model bears resemblance to Cruse's 1995 framework, which separates genuine meanings, termed "senses," from what he calls "facets," suggesting they are more relevant to reading and interpretation.

When applying Cruse's terminology to pun analysis, it becomes apparent that puns rely on fully developed senses to create effective semantic contrast. Unlike facets, which can coexist within a single context, senses display "mutual antagonism," where their individual presence in a context serves to clarify their meaning. While facets may only result in vagueness due to a lack of semantic distance necessary for pun creation, senses produce genuine ambiguity, a crucial element for most types of puns to manifest [10].


HOMONYMY
POLYSEMY
MONOSEMY
PUN
NON-PUN
→ (etymologically) *un*related *senses*
→ upper limit of semantic distance
→ ambiguity
→ (etymologically) related *senses*
→ lower limit of semantic distance
→ ambiguity
→ *facets* ('local senses')
→ *insufficient* semantic distance
→ vagueness

* Downward-pointing arrows mark the descending order of semantic distance.

Figure 1: Degree of Semantic Contrast in Puns vs. Non-puns

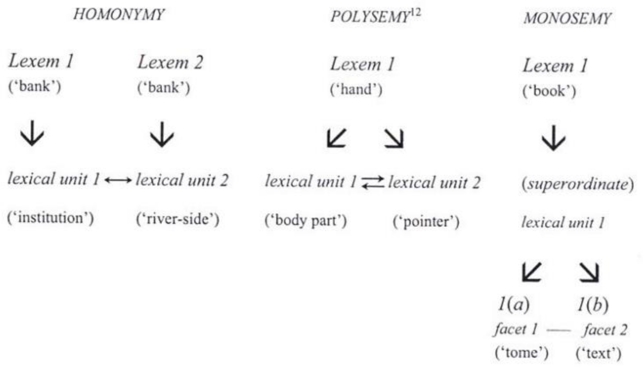


Figure 2: Examples for the Semantic Contrast

Semantic diversity is a metric that quantifies the level of ambiguity in word usage by examining the variability of contexts in which words appear. The ambiguity of a word arises when it has more than one sense or meaning. However, we need to investigate what degree of difference between two uses of a word, qualifies them as distinct senses.

Another approach to measuring semantic ambiguity moves away from the idea that words possess a fixed number of distinct senses or meanings. Instead, it acknowledges that each word carries a unique semantic "flavor" that can change depending on the context in which it is used. This variability in contextual interpretation, such as between "perjury" and "predicament," is not fully captured by the traditional definition of semantic ambiguity.

Contextual distinctiveness measures the predictability of a word's immediate context. In [7] they analyzed the distribution of words for each appearance of a particular word in their corpus. The LSA method utilizes a large corpus divided into several discrete contexts, where each context is a sample of text from a particular source. LSA as per their paper, then tabulates a cooccurrence matrix registering which words appear in which contexts.

Data-reduction technique called **Singular Value Decomposition** (SVD) is the crucial step reveals latent higher-order relationships between words, based on their patterns of co-occurrence [7]. In SemD, the researchers evaluated this for all contexts in which the word appears and took an average of this. More similarity in the contexts where a word appears suggests, according to them, that the word was associated with a more restricted set of meanings. This means it's relatively unambiguous. When the contexts associated with a given word were quite dissimilar to one another, this suggested that the meaning of the word was more [7].

Prior to SVD, values in the matrix were log-transformed. The logs associated with each word were then divided by that word's entropy (H) in the corpus:

$$H = \sum_{c} p_c log(p_c)$$

where c indexes the different contexts in which the word appears, and p_c denotes the word's frequency in the context divided by its total frequency in the corpus. These standard transformations were performed to reduce the influence of very high-frequency function words whose patterns of occurrence were not relevant in generating the semantic space [7]. SVD was then used to produce a solution with as many dimensions as there were contexts. The result of this process was two sets of vectors. First, there was a vector for each word in the corpus, describing its location in the semantic space. These are the vectors typically used in applications of LSA; similarity in the vectors of two words is thought to indicate similarity in their meanings. Second, there was a vector for each context that we analyzed, describing its location in the semantic space. The authors of the paper hypothesized that the similarity between the vectors of two contexts would indicate their similarity in semantic content. These context vectors were used in the calculation of SemD.

The authors also acknowledged shortcomings of their methodology such as the clustering of the contexts into different meanings. They use the example of homonyms like 'bark' which may occur primarily in two contexts- one where dogs bark , the other being the bark of a tree. There can be a high diversity between these two clusters of meaning but low diversity within each cluster. But many polysemous words may occur in a broad variety of contexts that do not cluster easily. In SemD, both will register high values, and may be construed to have similar diversity. They are not.

METHODOLOGY

The initial phase in grasping the contextual shifts within puns involves addressing the semantic variety of words within any given context by measuring this diversity.

We aim to accomplish this by employing the SemD technique [7] on collections of texts, serving as the preliminary stage in our exploration of polysemy and deliberate contextual shifts in puns more broadly.

I. DATASET DESCRIPTION AND PREPROCESSING

The text data utilized in this study was sourced from the Gutenberg-Dammit database [8], an archive that aggregates content from Project Gutenberg [9], a digital library offering free access to a vast collection of literary works. Our primary aim is to conduct an evaluation and quantification of the semantic diversity exhibited by words across the oeuvres of notable authors. By comparing the patterns of semantic shifts employed by these authors, we aim to gain insights into their linguistic styles and approaches to conveying meaning.

Our selection of authors was carefully curated, considering both the quantity and diversity of their literary output. From the timeless works of William Shakespeare to the thrilling mysteries penned by Arthur Conan Doyle, and from the witty observations of Gilbert Keith Chesterton to the profound insights of Leo Tolstoy, our selection encompasses a rich tapestry of literary talent. By including translated works of Leo Tolstoy, we further broaden the scope of our analysis, incorporating the nuances of semantic diversity across diverse cultural contexts.

The text undergoes preprocessing, where it is transformed into lowercase words and segmented into context chunks comprising thousand words each. To streamline the evaluation process, only words that have appeared at least five times and across a minimum of two contexts are selected for further analysis. This selection criterion ensures that the words chosen for evaluation exhibit a sufficient degree of variability in their contextual usage, allowing for a more nuanced assessment of semantic diversity.

II. MODEL CONSTRUCTION AND EVALUATION

Pseudo-code for calculating SemD based on [7]	
1.	Divide each document in the Corpus into “contexts” of thousand words.
2.	Generate a Co-Occurrence Matrix representing words and the contexts. The value at *(i, j)* represents the number of times word *i* appears in context *j*.
3.	Process the Co-occurrence matrix to normalize it against the entropy of the words. This is to ensure that frequently-used words – which are bound to be in more contexts- are corrected for their information density using word entropy values.
4.	Perform a Single Value Decomposition (SVD) of the matrix to get word vectors and context vectors.
5.	Using the LSA vectors for contexts, compute the similarity of all pairwise combinations of contexts containing the word by taking the cosine of their vectors.
6.	Take the mean of the cosines, then take the natural log of this and reverse the sign to give the semantic diversity value for the word.

While this is the intended methodology, since the paper does not explicitly provide the encryption to calculate SemD for a given corpus, we are required to decipher and reconstruct the code by ourselves, generating co-occurrence matrices and conducting latent semantic analysis and SemD calculations by ourselves, without using existing packages in Python. This has led to a few errors in calculation which we have yet to decipher and correct.

RESULTS

In Figure 3, we observe the distribution of semantic diversity within the corpus of William Shakespeare, highlighting the range of meanings and contexts in which words are employed throughout his works. Figure 4 provides a similar analysis for the corpus of Gilbert Keith Chesterton, offering a comparative view of semantic diversity within his literary compositions. Figure 5 illustrates the distribution of semantic diversity within the corpus of Oscar Wilde, providing further insights into the richness and complexity of his writing style. In Figure 6, we present the distribution of semantic diversity within the corpus of Arthur Conan Doyle, allowing for a comprehensive examination of the varied meanings and contexts present in his works. Figure 7 extends this analysis to the corpus of Mark Twain, offering additional insights into the semantic diversity within his literary creations. Figure 8 further enriches our understanding by illustrating the distribution of semantic diversity within the corpus of Rudyard Kipling, providing a comparative perspective on the usage of words in his writing style.

Lower SemD values imply lower semantic diversity. Higher SemD values imply higher semantic diversity.

As can be observed from the figures, certain words score high in SemD, these are words that can be construed as “common” words likelier to be used in a variety of ways and situations. The words on the right side of the plot can be understood to be words only used in very specific contexts, where the variation of context vectors scores low. However, these results are only preliminary and have several questions yet to be answered and flaws yet to be understood.

For example, we need to equate SemD values with the commonness of words and their information entropy values: the current “normalization” procedure does not seem to have had the desired effect of downplaying words simply for being common.

The size of contexts of thousand words is arbitrary. Depending on what we need to know, this context might be too big to successfully capture context specifics, and yet too small to genuinely encompass the scope of certain other contexts. Contexts are made of varying shapes and sizes.

CONCLUSION

The ongoing refinement of SemD calculations presents an opportunity for enhancing our understanding of semantic diversity within textual corpora. A key focus of this refinement is the normalization of words by their information values, a critical step in ensuring the accuracy and reliability of our semantic diversity assessments. To gain further insights into our SemD calculations and validate their effectiveness, we intend to compare them with established lexical databases such as WordNet. These databases categorize words based on various linguistic criteria, including their functions, phonological properties, and morphological characteristics. By juxtaposing SemD values with the classifications provided by WordNet, we can assess the extent to which our calculations accurately capture semantic nuances and distinctions within the text.

A crucial aspect of this comparative analysis involves examining the SemD values of typical pun words—such as homographs, homophones, and paronyms—against those of non-pun words. This comparative approach enables us to discern whether SemD values effectively differentiate between words used for comedic wordplay and those employed in straightforward linguistic contexts.

Finally, once we are confident in the accuracy and reliability of our SemD calculations, we can begin to meaningfully characterize an author's work or the stylistic characteristics of a particular genre based on semantic diversity. By quantifying and analyzing semantic variability within textual corpora, we gain deeper insights into the intricate dynamics of language usage and interpretation, paving the way for a more nuanced understanding of literary creativity and expression.

FUTURE WORK

Once we have refined our SemD measure, we aim to explore context in a more sophisticated manner. This involves employing advanced tools like transformers and other NLP techniques that capture context in a dynamic way, rather than relying on fixed 1000-word chunks as we currently do. Analyzing contexts through sentence vectors, paragraph vectors, and document vectors within domain-specific corpora will provide deeper insights into our tool.

We plan to compare semantic diversity values between puns and non-puns, homographs, and homophones, and examine how a word's SemD values vary across different corpora. For instance, we're interested in observing how the word "bank" fluctuates across various fields such as science, literature, and pop culture. This comparative analysis will shed light on the semantic nuances and usage patterns of words across diverse contexts.

“It has also long been argued that abstract words have inherently more variable and context-dependent meanings than do concrete or highly imageable words” [Hoffman et al., 2013, p. 722]

We want to examine how SemD values stack up against other abstraction measures to assess our word usage. Do abstract words fare better in various contexts? Does the broadness of abstractions result in greater semantic variety? The exploration of polysemy and its applications is vast and challenging. SemD is just one among numerous tools to help navigate this complexity.

REFERENCES

1. J. Lyons, Introduction to Theoretical Linguistics. Cambridge, UK: Cambridge Univ. Press, 1968.

2. S. Trott and B. Bergen, "Why do human languages have homophones?," Cognition, vol. 205, pp. 104449, 2020.

3. S. Valera and A. E. Ruz, "Conversion in English: homonymy, polysemy and paronymy," Eng. Lang. & Ling., vol. 25, no. 1, pp. 181-204, 2021.

4. D. Bergeron, "Heteronyms," Eng. Today, vol. 6, no. 4, pp. 39-44, 1990.

5. G. N. Leech and Internet Archive, Semantics. Harmondsworth, UK: Penguin, 1974. [Online]. Available: https://archive.org/details/semantics00leec

6. Y. Diao et al., "Homographic Puns Recognition Based on Latent Semantic Structures," in NLPCC 2017, X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, Eds. Cham: Springer, 2018, pp. 47. https://doi.org/10.1007/978-3-319-73618-1_47

7. P. Hoffman, M. A. Lambon Ralph, and T. T. Rogers, "Semantic Diversity: A Measure Of Semantic Ambiguity Based On Variability In The Contextual Usage Of Words," Behav. Res. Methods, vol. 45, pp. 718-730, 2013.

8. A. Parrish and H. van Kemenade, "aparrish/gutenberg-dammit," May 12, 2024. [Online]. Available: https://github.com/aparrish/gutenberg-dammit

9. Project Gutenberg, "Project Gutenberg." [Online]. Available: http://www.gutenberg.org. [Accessed: April 12, 2024].

10. W. Croft and D. A. Cruse, Cognitive Linguistics. Cambridge, UK: Cambridge Univ. Press, 2004.

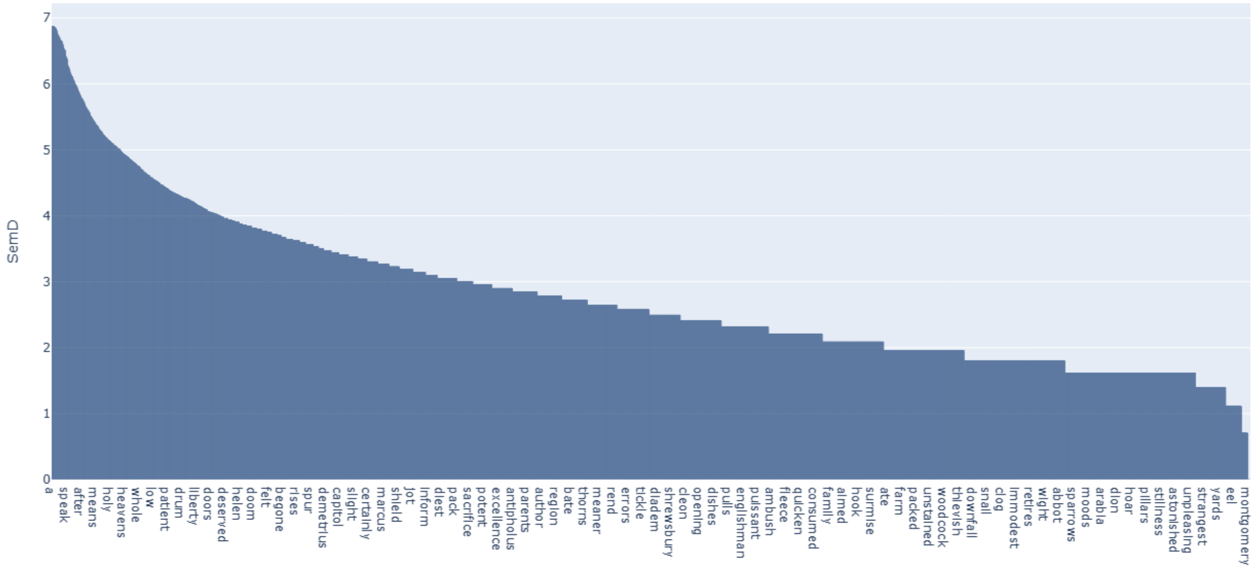


Figure 3. Semantic Diversity for Words in the Shakespeare Corpus

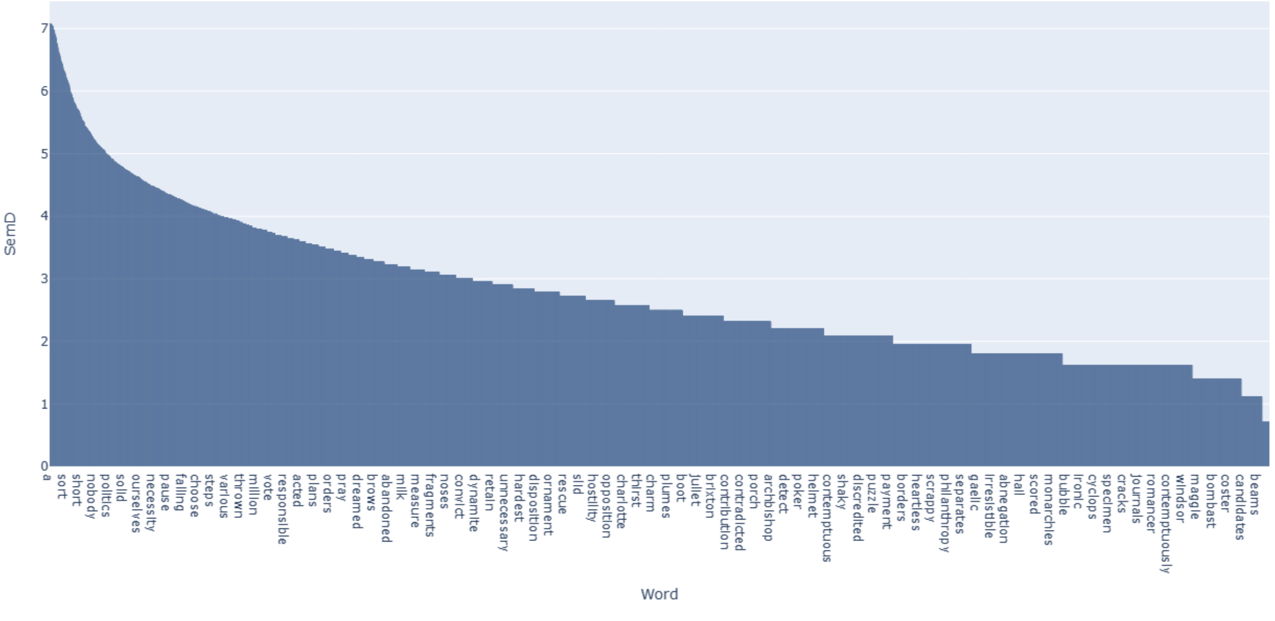


Figure 4. Semantic Diversity of Words in the Gilbert Keith Chesterton Corpus

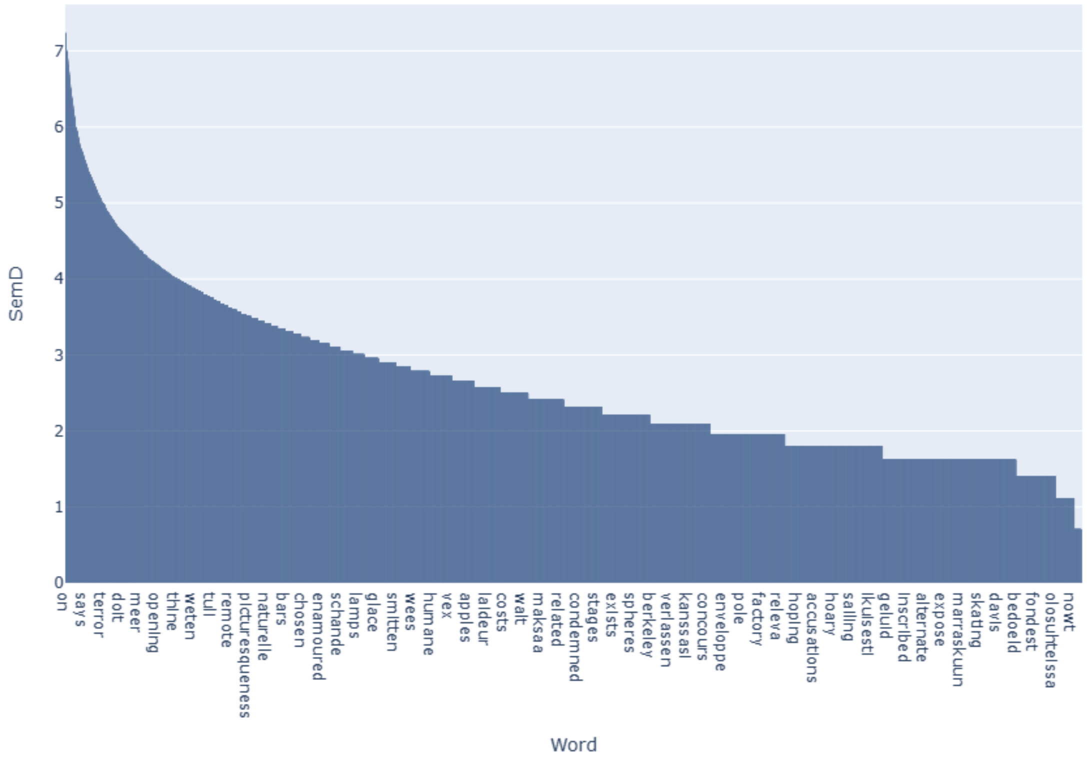


Figure 5. Semantic Diversity of Words in the Oscar Wilde Corpus

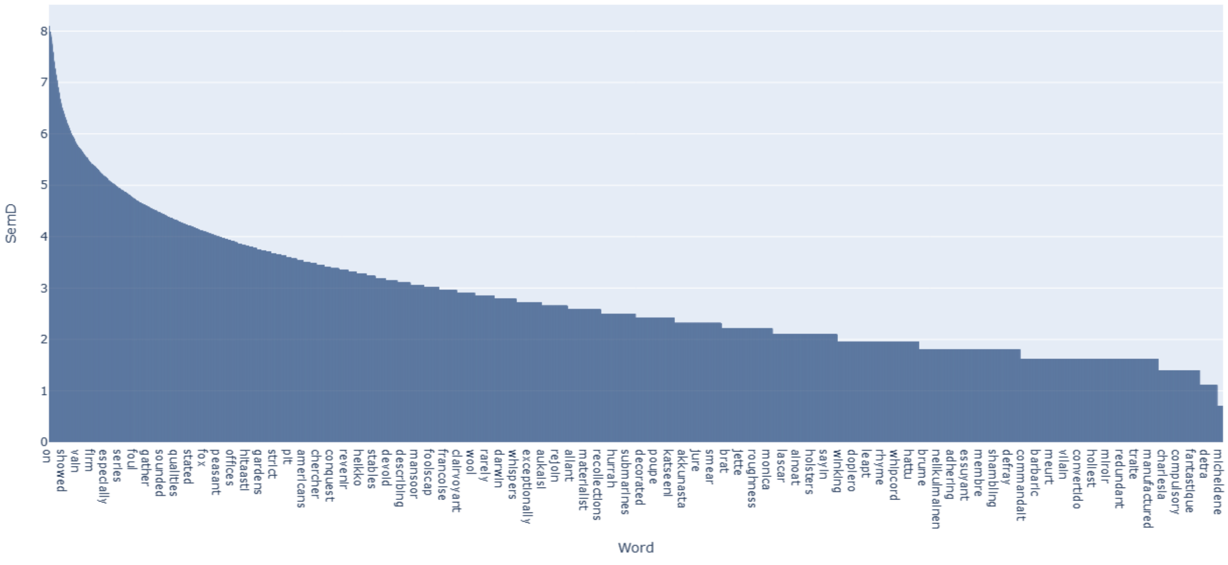


Figure 6. Semantic Diversity for Words in Arthur Conan Doyle Corpus

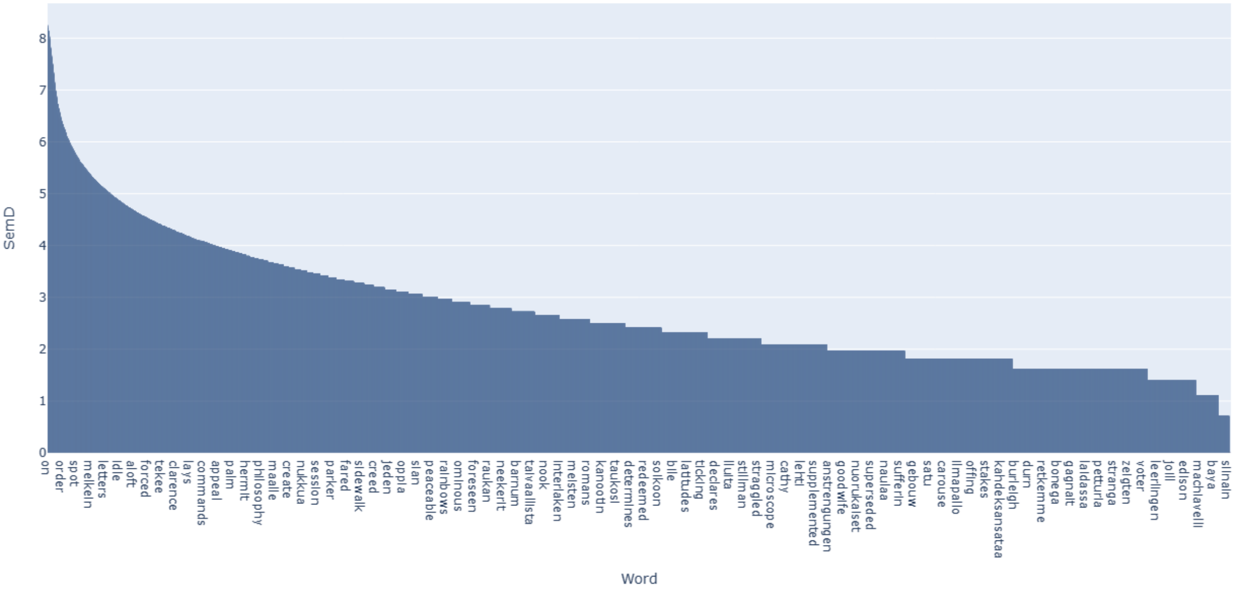


Figure 7. Semantic Diversity of Words in Mark Twain Corpus

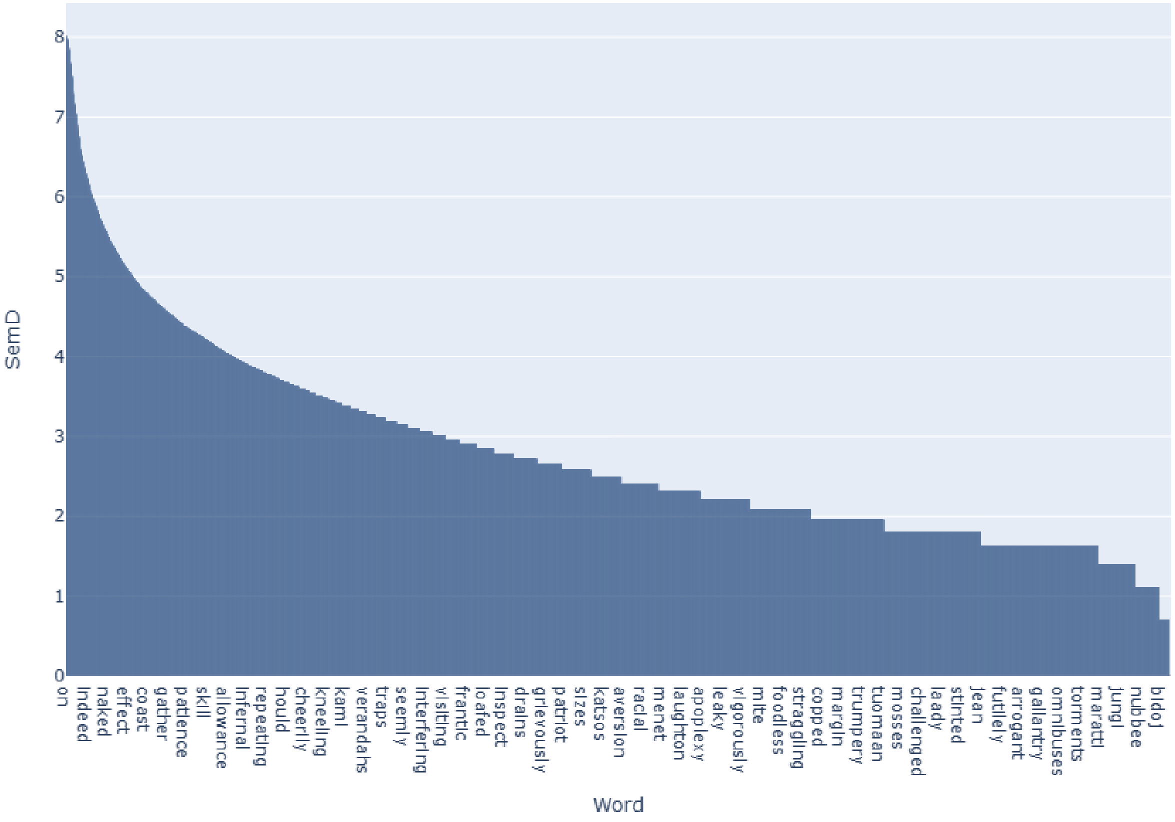


Figure 8. Semantic Diversity of Words in Rudyard Kiliping Corpus