# SentSpace: Large-Scale Benchmarking and Evaluation of Text using Cognitively Motivated Lexical, Syntactic, and Semantic Features

**Greta Tuckute*** **Aalok Sathe*** **Mingye Wang◇** **Harley Yoder◇**
**Cory Shain** **Evelina Fedorenko**

{gretatu,asathe,mingyew,hyoder,cshain,evelina9} @ mit.edu

Dept. of Brain and Cognitive Sciences McGovern Institute for Brain Research
Massachusetts Institute of Technology, Cambridge, MA, USA

## Abstract

SentSpace is a modular framework for streamlined evaluation of text. SentSpace characterizes textual input using diverse lexical, syntactic, and semantic features derived from corpora and psycholinguistic experiments. Core sentence features fall into three primary feature spaces: 1) *Lexical*, 2) *Contextual*, and 3) *Embeddings*. To aid in the analysis of computed features, SentSpace provides a web interface for interactive visualization and comparison with text from large corpora. The modular design of SentSpace allows researchers to easily integrate their own feature computation into the pipeline while benefiting from a common framework for evaluation and visualization. In this manuscript we will describe the design of SentSpace, its core feature spaces, and demonstrate an example use case by comparing human-written and machine-generated (GPT2-XL) sentences to each other. We find that while GPT2-XL-generated text appears fluent at the surface level, psycholinguistic norms and measures of syntactic processing reveal key differences between text produced by humans and machines. Thus, SentSpace provides a broad set of cognitively motivated linguistic features for evaluation of text within natural language processing, cognitive science, as well as the social sciences.

## 1 Introduction

Natural Language Processing (NLP) researchers and language scientists alike rely heavily on numeric representations of text in order to better understand how machines and humans process language. Consider the following text generated by a large pre-trained language model, GPT2:

> The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

The passage demonstrates a remarkable facility with language, but also some potentially non-human aspects, both syntactic (e.g., an unnatural *forward shifting* of the phrase *"after their distinctive horn"*) and semantic (describing a four-horned animal as a unicorn). It is of growing interest to researchers to be able to characterize how any text compares to that from another source, be it generated by humans or artificial language models. For example, NLP practitioners are interested in understanding and improving language model output, and bringing it closer to human generated text e.g., Ettinger (2020); Hollenstein et al. (2021); Meister et al. (2022); similarly, language scientists are highly interested in using large-scale language models to develop and test hypotheses about language processing in the mind and brain, e.g., Schrimpf et al. (2021); Caucheteux and King (2022); Goldstein et al. (2022).

To support these shared goals, we developed SentSpace, an open source application for characterizing textual input using diverse lexical, syntactic, and semantic features. These features are derived from sources such as large, constructed corpora, behavioral psycholinguistic experiments, human judgment norms, and models based on theories of human sentence processing. We also developed functionality to compare textual inputs to one another and to large normative distributions based on natural language corpora. We envision the use cases of SentSpace to be diverse: **(i)** comparison of machine-generated text to human-generated text; **(ii)** comparison of text produced by different human populations (e.g., native and non-native speakers, neurotypical individuals and individuals with developmental or acquired communication disorders); **(iii)** comparison of different genres of text; **(iv)** evaluation of the normativity of stimuli/datasets to be used in psycholinguistic experiments or experiments with language models; and **(v)** investigation of sentences that present particular comprehension

*Equal contribution ◇Equal contribution

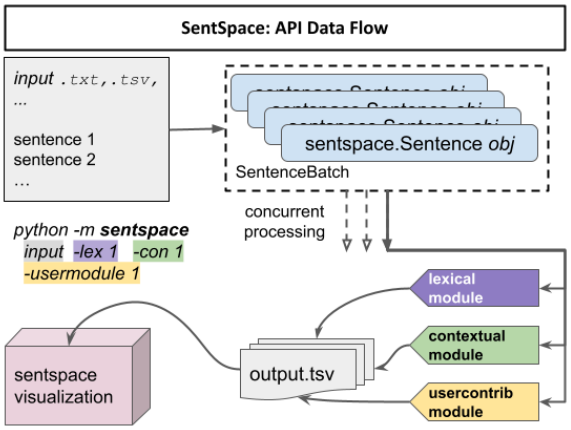


Figure 1: An overview of `SentSpace` data flow: users can supply text in a number of formats. The text is batch-processed in each selected module containing various features. The features computed from each module are then outputted in the specified output file format. These results can be readily plugged into the visualization module out of the box.

difficulties for humans and/or language models, e.g., eliciting a strong neural response in an electrophysiological or neuroimaging study, or producing non-typical or undesired behavior in the case of language models.

We make `SentSpace` publicly available via https://sentspace.github.io/sentspace in the following two forms: **(i)** an open source API implemented in Python (Figure 1), installable via the Python package index (PyPI) or as a self-contained Docker image; **(ii)** a hosted web interface for computing and visualizing features, thus making `SentSpace` accessible without running locally (Figures 2, 3, 4).

## 2 Structure and Design

At the core of `SentSpace` there are features associated with a sentence: $f$ : `sentence` $\rightarrow \mathbb{R}^d$ where $\mathbb{R}^d$ can be any feature or representation space. The core features are organized into three core modules based on the nature of their characterization of a sentence. **(1)** The *Lexical* module acts at the individual lexical item (token) level. Sentence-level lexical features are computed by aggregating over the tokens of a sentence. **(2)** *Contextual* features are sentence-level features and are obtained as a result of some computation at the sentence level, such as, constructing and then processing a syntax parse tree. Finally, **(3)** *Embeddings* computes pooled vector representations from one of many popular embedding models, from GloVE (Pennington et al., 2014) to Transformer architectures (Radford et al., 2019; Devlin et al., 2019; Wolf et al., 2020). These three modules cover a wide range of features—derived from text corpora or behavioral experiments—that have some demonstrated relevance to language processing.

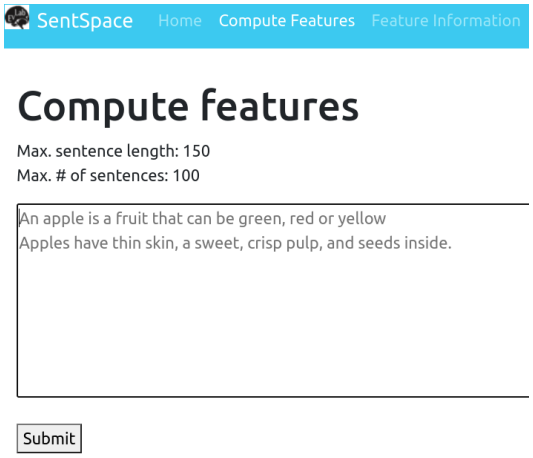


Figure 2: Public-facing hosted interface where users can input text and obtain features by downloading them from the same website. Each request gets a corresponding ID which is temporarily cached to enable repeated downloading and visualization of the same data.

The *Lexical* module consists of features that pertain to individual lexical items, words, regardless of the context in which they appear. These features include, for example, lexical frequency, concreteness, and valence. Because `SentSpace` is built to work with sentences, lexical-level features are aggregated across the sentence (cf. Gao et al. (2021)). As a default, `SentSpace` aggregates over all words with available norms in the sentence by computing the arithmetic mean across words.

The *Contextual* module also consists of features that quantify contextual and combinatorial inter-word relations that are not captured by individual lexical items. This module encompasses features that relate to the syntactic structure of the sentence (*Contextual_syntax* features) and features that apply to the sentence context but are not (exclusively) related to syntactic structure (*Contextual_misc* features). *Contextual_syntax* include features related to syntactic complexity, instantiated as e.g., surprisal or integration cost, based on leading theoretical proposals (Gibson, 2000; Shain et al., 2016; Rasmussen and Schuler, 2018). Some syntactic features are computed for each word in the sentence and then subsequently aggregated; other features are computed for multi-word sequences or the entire sentence. *Contextual_misc* include features

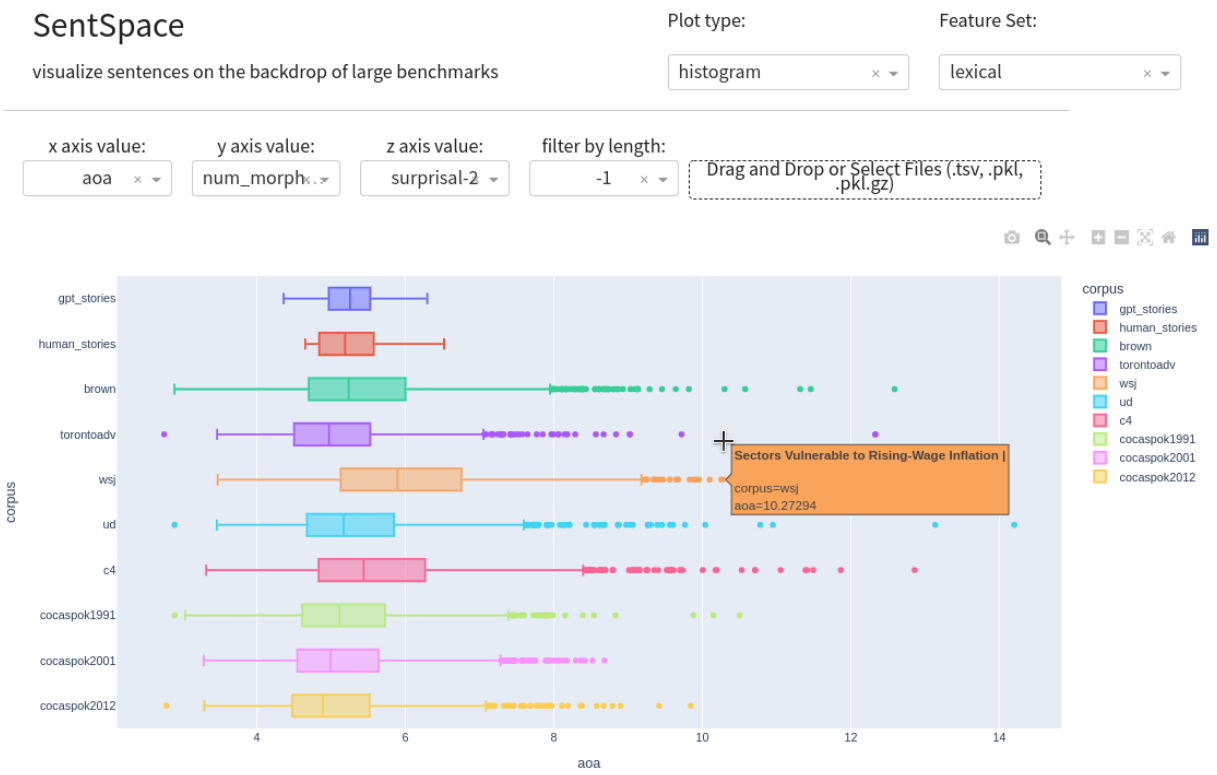


Figure 3: An example of multiple corpora visualized alongside each other for the feature "Age of Acquisition". Sentences from the Wall Street Journal and the Colossal Cleaned Common Crawl (C4) show a tendency of higher age of acquisition on average than other sources. Mouseover on dots enables users to see example sentences and their corresponding values. The 'x axis value' dropdown allows users to pick the feature to plot. The 'y' and 'z' axis values are used for a 3D scatter plot, which can be enabled using the 'Plot type' selector.

like lexical density or sentence sentiment.

The *Embedding* module consists of high-dimensional vectors derived from pre-trained language models.

`SentSpace` also provides functionality that allows users to contribute novel features and modules. A user may design their own features and plug-and-play into `SentSpace` to achieve a more streamlined analysis pipeline and integrated benchmarking and visualization (Figure 1). In order to contribute a module, users must adhere to the module call API, accepting a sentence batch and returning a dataframe whose columns consist of features. Users may make use of parallelism and other utils provided as a part of `SentSpace`. Users may also plug in their computed features in the visualization module and use the web interface.

## 2.1 Feature Modules

### 2.1.1 Lexical

Lexical features have been shown to affect language comprehension at the level of individual words. For instance, lexical features affect how people recognize and recall words, such as word frequency (e.g., Gorman (1961); Kinsbourne and George (1974)), concreteness/imageability (e.g., Gorman (1961); Rubin and Friendly (1986)), and valence/arousal (e.g., (Rubin and Friendly, 1986; Danion et al., 1995; Kensinger and Corkin, 2003). Moreover, lexical features have been shown to affect language processing when words are presented in context as measured by eye tracking and self-paced reading, such as surprisal (e.g., Levy (2015); Demberg and Keller (2008); Singh et al. (2016)), polysemy (e.g., Pickering and Frisson (2001)), ambiguity (e.g., Frazier and Rayner (1987); Rayner and Duffy (1986)), word frequency (e.g., Rayner and Duffy (1986)), and age of acquisition (e.g., Singh et al. (2016)). We implement these features using lookup tables for each token. In case the feature is unavailable for a token, we use a lemmatizer to obtain the feature corresponding to the word's lemma. We observe the various features are only moderately correlated with one another, thus each adding new information to the analysis (Figure 5). See Appendix A.1 for supported lexical features.

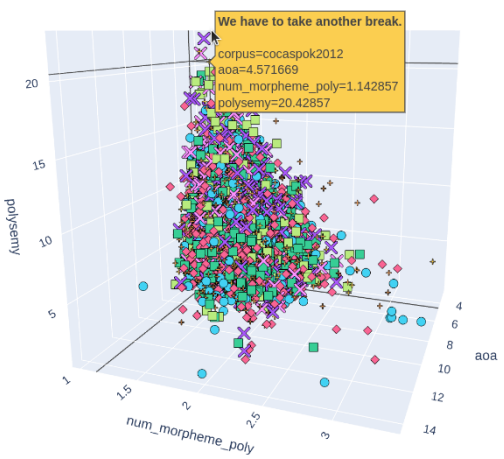


Figure 4: A zoomed-in view of a 3D scatterplot shows mouseover on a point in space revealing the sentence at that location and its features being plotted. A top bar (not displayed) allows the users to change the features to plot. A side bar (not displayed) enables selecting/deselecting corpora and files uploaded by the user to the visualization module. (Abbreviations are 'num_morpheme_poly': Number of Morphemes; 'aoa': Age of Acquisition.)

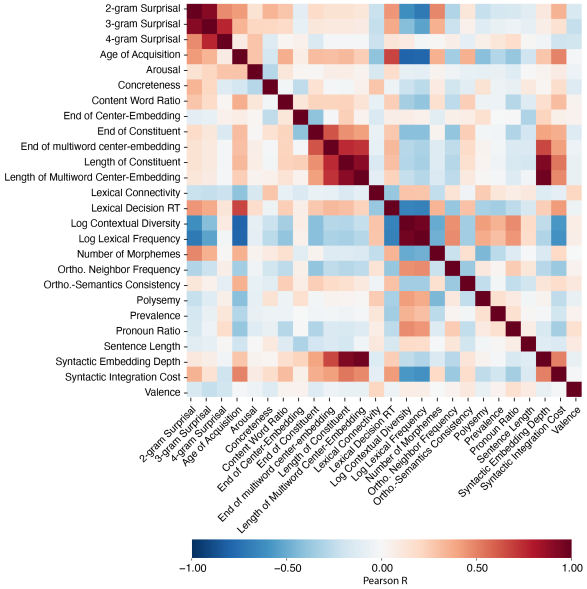


Figure 5: Pearson correlation among features from the *Lexical* and *Contextual* modules obtained from `SentSpace` for text written by humans and GPT2-XL (described in Section 4).

### 2.1.2 Contextual

Several properties of a sentence cannot be attributed to its individual lexical items (words). These features broadly fall into two categories: syntactic (denoted by *Contextual_syntax*) and miscellaneous (denoted by *Contextual_misc*). The syntactic features include measures of storage and integration cost as predicted by both the Dependency Locality Theory (DLT; Gibson (2000)) and left-corner theories of sentence processing (Rasmussen and Schuler, 2018). In brief, the Dependency Locality Theory is an influential theory of word-by-word comprehension difficulty during human language processing, with difficulty hypothesized to arise from working memory demand related to storing items in working memory (storage cost) and retrieving items from working memory (integration cost) as required by the dependency structure of the sentence. Memory costs derived from the DLT have been associated with self-paced reading (Grodner and Gibson, 2005), eye-tracking (Demberg and Keller, 2008), and fMRI (Shain et al., 2021b) measures of comprehension difficulty. Left-corner parsing models also posit storage and integration costs, but these costs are thought to derive not from dependency locality but from the number of unconnected fragments of phrase structure trees that must be maintained and combined in memory throughout parsing, word-by-word. Probabilistic left-corner parsers can also be used to define a probability distribution over the next word that conditions solely on hypotheses about the syntactic structure of the sentence, providing a critical tool for evaluating the degree to which syntax might influence both human and language model predictions of future words (Shain et al., 2020). See Appendix A.2.1 for supported contextual features.

### 2.1.3 Embeddings

Embeddings provide representations of words or sentences in high-dimensional, learned vector spaces. The information contained in these spaces depend on the objective function of the algorithm used to derive the vectors, but could be of semantic nature (e.g., Grand et al. (2022)). We provide a decontextualized embedding space (words have the same vector representation independent of context), GloVe (Pennington et al., 2014), as well as several commonly used contextualized embedding spaces (words have different vector representations based on the context in which they appear) from the HuggingFace framework (Wolf et al., 2020). See Appendix A.3 for supported embedding models.

## 3 Benchmarking Against Large Corpora

To understand where a sentence stands relative to other text, we facilitate comparison with sentences

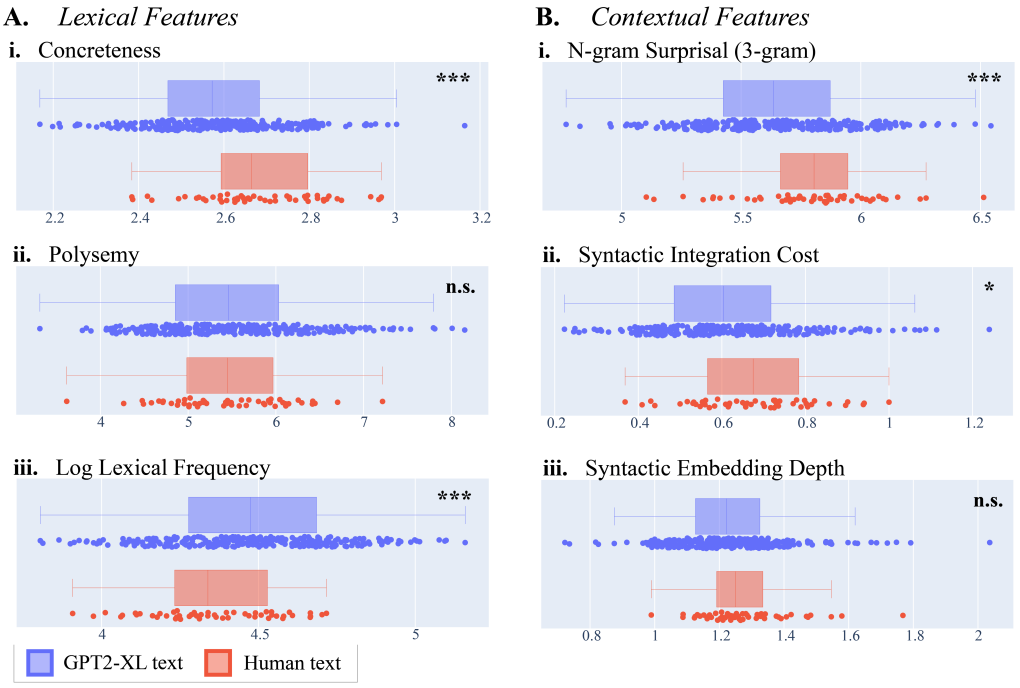


Figure 6: We use `SentSpace` to visualize sentences from two sources of interest. In one case, humans generated paragraphs, and in the other, a GPT2-XL language model did. We find several points of differences between the two sources, verified using statistical tests comparing the two distributions. $p$-values were obtained using two-tailed independent samples $t$-tests: Concreteness ($t = 4.24, p \ll 0.001$), Polysemy ($t = -0.27, p = n.s.$), Lexical Frequency ($t = -2.91, p < 0.005$), N-gram Surprisal (3-gram) ($t = 2.91, p < 0.005$), Syntactic Integration Cost ($t = 2.34, p < 0.05$), Syntactic Embedding Depth ($t = 1.81, p = n.s.$). $^*p < .05$, $^{**}p < .01$, $^{***}p < .005$.

from large corpora of human-generated text (both written and spoken). We allow this by subsampling from large corpora to include an approximately equal number ($\approx 500$) of sentences from each corpus. We pre-computed and cached `SentSpace` features for each of 8 corpora ($\approx 4000$ sentences in total), enabling quick and streamlined comparison with sentences from existing corpora. In the `SentSpace[.vis]` module, these corpora are loaded by default in addition to user-supplied input. Corpora benchmarking can be disabled to allow visualizing user input in isolation. We provide a list of corpora used in Appendix C.

## 4 System Demonstration and Results

In this section we provide an example of how `SentSpace` can be used to compare and visualize sets of sentences to one another using features from the *Lexical* ($n = 13$ features) and *Contextual* ($n = 13$ features) modules of `SentSpace`[1]. For this example demonstration, we compare two sets of materials: Human-generated text versus GPT2-XL-generated text. The texts consisted of 52 unique paragraphs written by multiple human writers. The first 10 words of each paragraph were used as a prompt to a pretrained GPT2-XL autoregressive language model (Radford et al., 2019; Wolf et al., 2020). Prompt completions were extracted across multiple random seeds using top-p sampling (Holtzman et al., 2020) with generation parameters $p = 0.9$ and $temperature = 1$. We selected 5 completions per prompt that most closely matched the human-generated prompt in word length (within $\pm 5$ words) to control for any length-driven correlations. As a result, we had one human-generated and 5 GPT2-XL-generated paragraphs per prompt, yielding a total of $n = 52$ human-generated paragraphs and $n = 260$ GPT2-XL-generated paragraphs (for examples, see Appendix B). Features were averaged across sentences within each paragraph. For statistical tests, features for the $n = 5$ GPT2-XL-generated paragraphs for the same prompt were averaged to yield a matched sample of paragraphs with the human-generated paragraphs ($n = 52$). In Figure 6, we demonstrate that our feature measures can reveal subtle quantitative differences between machine-generated (blue) and human-generated (red) texts that may not be

[1]The code for analyses in this paper is available at http://github.com/sentspace/NAACL-HLT-2022

subjectively apparent.

Figure 6A demonstrates three features from the *Lexical* module: **i)** Concreteness (Brysbaert et al., 2014); a behavioral measure of the extent to which the concept denoted by the word refers to a perceptible entity, **ii)** Polysemy (Miller, 1992), and **iii)** Log lexical frequency from the SUBTLEX-us database (Brysbaert and New, 2009). As evident, GPT2-XL produces sentences that on average have less concrete words compared to human sentences ($p \ll .001$). Lexical frequency reflects how often a given word is used in language. Lexical frequency is known to affect language comprehension, for instance more frequent words are read faster (e.g., Rayner and Duffy (1986); Singh et al. (2016) and articulated faster (e.g., Jescheniak and Levelt (1994)). We can see this as being a trend towards GPT2-XL's use of more frequent wording compared to humans ($p \ll .001$).

Figure 6B demonstrates three *Contextual* features: **i)** N-gram surprisal (3-gram), **ii)** Average syntactic integration cost according to the Dependency Locality Theory (DLT, (Gibson, 2000); integration cost is roughly proportional to dependency length), and **iii)** Average syntactic center-embedding depth in a left-corner phrase-structure parser (van Schijndel et al., 2013). Although GPT2-XL usually generates sentences that are syntactically well-formed, their syntactic features differ on average from human-generated text. As shown, texts generated by GPT2-XL show lower 3-gram surprisal ($t$=2.91, $p \ll .001$), tend to be less syntactically complex on average than human-generated ones, with shorter syntactic dependencies ($t$=2.33, $p$=0.02) and numerically shallower center-embedded tree structures ($t$=1.8, $p$=0.09, n.s.). So, these findings might suggest GPT2-XL makes use of 'simpler' wording compared to humans.

The remaining `SentSpace` features obtained for the comparison between human- and GPT2-XL-generated text ($n = 26$ features in total) are summarized in the Appendix, Table 2. More features are in the progress of being added to the `SentSpace` framework (see Appendix A.1, A.2).

The comparison between human- and machine-generated text is a demonstration of one of the use cases of `SentSpace`: comparing and visualizing texts to one another. The `SentSpace` framework streamlines the process of obtaining corpora-backed features, parsing and syntactically analyzing texts, simplifying and accelerating such analyses for natural language generation.

## 5 Related work

Related work include Balota et al. (2007) who collected behavioral visual lexical decision and speeded naming reaction times and provided these along with a set of word-level, psycholinguistic features (The English Lexicon Project). Gao et al. (2021) provide a meta-base of word-level, psycholinguistic features. A different alley of related work includes visualization tools for high-dimensional embeddings obtained from pre-trained language models (e.g., van Aken et al. (2020); OpenAI).

A large body of work focuses on characterizing bias in text, particularly that either used in training language models, or that generated by language models (Sun et al., 2019). Related work also focuses on methods to mitigate bias in existing language models using debiasing methods. In the future we hope to include norms that characterize bias as one of the many features that will be added to `SentSpace`. We also hope that outputs from `SentSpace` will inform what data goes into training large language models to make them more human-like.

## 6 Conclusion

`SentSpace` is a system for obtaining numerical representations of sentences. Our core feature modules span lexical, semantic, and syntactic features from corpora and behavioral experiments. We provide an interface for comparing textual inputs to one another or to large normative distributions based on natural language corpora.

Within the last few years, contextualized embeddings obtained from large pre-trained language models have revolutionized and dominated the field of natural language processing. However, despite these embeddings being useful for diverse applications, it is unclear precisely which information is embedded in these high-dimensional feature representations. We view `SentSpace` as a complementary resource that can provide interpretability and grounding to these pre-trained high-dimensional embeddings.

A major limitation of `SentSpace` is that we currently only support English. Part of the limiting factor is the relative lack of behavioral and psycholinguistic experimental data for other languages, as well as mature linguistic features tai-

lored to other languages.

We envision SentSpace as a dynamic platform with continuous collaboration across research labs for the addition of new features and we hope to make this framework valuable for a number of applications within natural language processing, cognitive science, psychology, linguistics, and social sciences.

## Acknowledgements

We thank the authors of publicly available datasets that we have been able to use in SentSpace. We thank Adil Amirov, Alvincé Le Arnz Pongos, Benjamin Lipkin, and Josef Affourtit for their assistance towards developing the software for SentSpace. We thank Hannah Small and Matthew Siegelman for their assistance with the human- and GPT-generated texts.

## References

David A. Balota, Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. 2007. The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Thorsten Brants and Alexander Franz. 2009. Web 1t 5-gram, 10 european languages version 1. *Philadelphia, Pa.: Linguistic Data Consortium*, Computer file.

Marc Brysbaert, Paweł Mandera, Samantha F. McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2):467–479.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, 44(4):991–997.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Charlotte Caucheteux and J. R. King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5.

J.-M. Danion, Françoise Kauffmann-Muller, Daniel le Grange, M A Zimmermann, and Ph. Greth. 1995. Affective valence of words, explicit and implicit memory in clinical depression. *Journal of affective disorders*, 34 3:227–34.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor*, 5(2):7.

Lyn Frazier and Keith Rayner. 1987. Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, 26(5):505–526.

Chuanji Gao, Svetlana V. Shinkareva, and Rutvik H. Desai. 2021. Scope: The south carolina psycholinguistic metabase.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain*, pages 95–106. MIT Press.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Rose Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner K. Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Y. Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25:369 – 380.

Anna M. Gorman. 1961. Recognition memory for nouns as a function of abstractness and frequency. *Journal of experimental psychology*, 61:23–9.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, 29 2:261–90.

Paul Hoffman, Matthew A. Lambon Ralph, and Timothy T. Rogers. 2013. Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, 45(3):718–730.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena A. Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *NAACL*.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Jörg D. Jescheniak and Willem J. M. Levelt. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20:824–843.

Elizabeth A. Kensinger and Suzanne Corkin. 2003. Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition*, 31:1169–1180.

Marcel Kinsbourne and James W. George. 1974. The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, 13:63–69.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30 thousand English words. page 39.

Roger Levy. 2015. Memory and surprisal in human sentence comprehension. page 36.

Matthew H. C. Mak and Hope Twitchell. 2020. Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired-associate learning. *Psychonomic Bulletin & Review*, 27(5):1059–1069.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguistics*, 19:313–330.

Marco Marelli and Simona Amenta. 2018. A database of orthography-semantics consistency (OSC) estimates for 15,017 English words. *Behavior Research Methods*, 50(4):1482–1495.

D.A. Medler and J.R. Binder. 2005. Mcword: An online orthographic database of the english language. volume http://www.neuro.mcw.edu/mcword/.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation. *ArXiv*, abs/2202.00666.

George A. Miller. 1992. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41.

Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *COLING*.

OpenAI. Openai embeddings. `https://beta.openai.com/docs/guides/embeddings`. Accessed: 2022-02-11.

Douglas B Paul and Janet Baker. 1992. The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

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)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Penny M. Pexman, Emiko Muraki, David M. Sidhu, Paul D. Siakaluk, and Melvin J. Yap. 2019. Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 English words. *Behavior Research Methods*, 51(2):453–466.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Martin J. Pickering and Steven Frisson. 2001. Processing ambiguous verbs: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(2):556–573.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nathan Rasmussen and William Schuler. 2018. Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive science*, 42 Suppl 4:1009–1042.

Keith Rayner and Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

David C. Rubin and Michael Friendly. 1986. Predicting which words get recalled: Measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns. *Memory & Cognition*, 14:79–94.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Cory Shain, Idan A Blank, Evelina Fedorenko, Edward Gibson, and William Schuler. 2021a. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *bioRxiv*.

Cory Shain, Idan Asher Blank, Evelina Fedorenko, Edward Gibson, and William Schuler. 2021b. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *bioRxiv*.

Cory Shain, Idan Asher Blank, Marten van Schijndel, Evelina Fedorenko, and William Schuler. 2020. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *CL4LC@COLING 2016*.

Abhinav Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *CL4LC@COLING 2016*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2020. Visbert: Hidden-state visualizations for transformers. *Companion Proceedings of the Web Conference 2020*.

Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Marten van Schijndel, Andrew Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in cognitive science*, 5 3:522–40.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

WordNet Wikipedia. Wordnet. `https://en.wikipedia.org/wiki/WordNet`. Accessed: 2022-02-11.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

# A Feature Descriptions

## A.1 Lexical Features

Lexical features are ordered alphabetically.

**Age of Acquisition** Age of acquisition is a metric that estimates when a person acquired (i.e., understood) a given word. Participants were asked for each word to enter the age at which they thought they had learned the word even if they could not use, read, or write it at the time. The norms were collected by (Kuperman et al., 2012).

Norms were available for 30,124 unique words.

Obtained from: http://crr.ugent.be/archives/806.

**Arousal** Arousal quantifies each word on a scale of active–passive. The norms were collected based on human ratings by (Mohammad, 2018) using Best-Worst scaling, where participants are presented with four words at a time and asked to select the word with the highest arousal. The two ends of the arousal dimension were: MOST arousal, activeness, stimulation, frenzy, jitteriness, alertness OR LEAST unarousal, passiveness, relaxation, calmness, sluggishness, dullness, sleepiness.

Norms were available for 20,007 unique words.

Obtained from: https://saifmohammad.com/WebPages/nrc-vad.html.

**Concreteness** Concreteness quantifies the extent to which the concept denoted by the word refers to a perceptible entity. Concrete words were defined as something that exists in reality and can be experienced directly through the five senses or actions. Conversely, abstract words refer to something one cannot experience directly through your senses or actions. The norms were collected based on human ratings by (Brysbaert et al., 2014).

Norms were available for 37,058 unique words.

Obtained from http://crr.ugent.be/archives/1330.

**Log Contextual Diversity** Contextual diversity (CD) is the number of contexts in which a word has been seen [cite Adelman 2005]. The metric available here was computed based on the SUBTLEXus database based on American subtitles (51 million words in total) (Brysbaert and New, 2009) and thus denotes the number of films in which the word appears. The CD metric is computed as log10(CDcount+1).

Norms were available for 74,286 unique words.

Obtained from the SUBTLEXus database: https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus).

**Log Lexical Frequency** Lexical frequency denotes how frequent a word occurs in a given corpus/sets of corpora. The frequency metric available here was computed as log10(FREQcount+1) based on American subtitles (51 million words in total) from the SUBTLEXus database (Brysbaert and New, 2009).

Norms were available for 74,286 unique words.

Obtained from the SUBTLEXus database: https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus).

**Lexical Connectivity** Lexical connectivity is a metric for how connected a given word is to other words based on association judgments. The metric views the mental lexicon as a semantic network where words are linked together by semantic relatedness. Lexical connectivity is computed as the total degree centrality of a given word node in the semantic graph. Norms were obtained from (Mak and Twitchell, 2020) who computed the total degree centrality based on free association norms collected by (De Deyne et al., 2019) (specifically, the first recalled word).

Norms were available for 12,215 unique words.

Obtained from: https://osf.io/7942s/.

**Lexical Decision Reaction Time (RT)** Lexical decision latency measures how quickly people classify strings as words or non-words. The lexical decision latency provides a proxy for how quickly a given word is extracted from the mental lexicon/semantic memory. The norms were collected by (Balota et al., 2007).

Norms were available for 40,482 unique words.

Obtained from the English Lexicon Project: https://elexicon.wustl.edu/.

**Number of Morphemes** A morpheme denotes the smallest meaningful lexical unit in a language. The number of morphemes quantifies how many morphemes a given word has. The primary morpheme counter available here is Morfessor (Virpioja et al., 2013) which uses machine learning to find morphological segmentations of words. If dependency issues arise with Morfessor, the morpheme count is obtained from the English Lexicon Project Database (Balota et al., 2007).

**Orthographic Neighbor Frequency** Orthographic neighbor frequency is a metric that quan-

tifies the number of orthographic neighbors that a string has. The metric was computed by (Medler and Binder, 2005) and an orthographic neighbor was defined as a word of the same length that differs from the original string by only one letter. The frequency metric denotes the averaged frequency (per million) of orthographic neighbors.

Norms were available for 66,371 unique words.

Obtained from http://www.neuro.mcw.edu/mcword/.

**Orthography-Semantics Consistency (OSC)** Orthography–semantics consistency is a metric that quantifies the degree of semantic relatedness between a word and other words that are orthographically similar. The metric was computed by (Marelli and Amenta, 2018) as the frequency-weighted average semantic similarity between the meaning of a given word and the meanings of all the words containing that very same orthographic string.

Norms were available for 15,017 unique words.

Obtained from: http://www.marcomarelli.net/resources/osc.

**Polysemy** Polysemy provides a metric of how many distinct meanings a word has. Polysemy was measured by the number of definitions of a word in WordNet (Miller, 1992). Polysemy was implemented using NLTK's word_net library (synsets() function) which accepts a word and a part-of-speech tag as input and returns a list of synonyms. Parts-of-speech tags were taken from NLTK's pos_tag, then mapped to the four POS tags accepted by word_net. If a POS tag could not be mapped to one of word_net's ADJ, VERB, NOUN, or ADV then the tag given was an empty string. The number of synonyms for a given word were counted.

Norms were available for 155,327 words organized in 175,979 synsets for a total of 207,016 word-sense pairs (Wikipedia).

Obtained from the NLTK interface: https://www.nltk.org/howto/wordnet.html.

**Prevalence** Word prevalence is a metric that quantifies the number of people who know a given word. The norms were collected by (Brysbaert et al., 2019) based on human ratings of whether or not they knew the word.

Norms were available for 61,855 unique words.

Obtained from: https://osf.io/g4xrt/.

**Valence** Valence quantifies each word on a scale of positiveness–negativeness. The norms were collected based on human ratings by (Mohammad, 2018) using Best-Worst scaling, where participants are presented with four words at a time and asked to select the word with the highest valence. The two ends of the valence dimension were: MOST happiness, pleasure, positiveness, satisfaction, contentedness, hopefulness OR LEAST unhappiness, annoyance, negativeness, dissatisfaction, melancholy, despair.

Norms were available for 20,007 unique words.

Obtained from: https://saifmohammad.com/WebPages/nrc-vad.html.

The following features were not analyzed in the current work, but in the future we plan to add support for these features in the `SentSpace` framework:

**Body-Object Interaction** Body-object interaction quantifies the ease with which the human body can interact with what a word represents. The norms were collected using behavioral ratings on a scale from 1 to 7 with a value of 7 indicating a high body-object interaction rating by (Pexman et al., 2019).

The norms were available for 9,349 unique words.

Obtained from: https://link.springer.com/article/10.3758%2Fs13428-018-1171-z#Sec9.

**Dominance** Dominance quantifies each word on a scale of dominant–submissive. The norms were collected based on human ratings by (Mohammad, 2018) using Best-Worst scaling, where participants are presented with four words at a time and asked to select the word with the highest dominance. The two ends of the dominance dimension were: MOST dominant, in control of the situation, powerful, influential, important, autonomous OR LEAST submissive, controlled by outside factors, weak, influenced, cared-for, guided.

Norms were available for 20,007 unique words.

Obtained from: https://saifmohammad.com/WebPages/nrc-vad.html.

**Part-of-Speech Ambiguity** Parts-of-speech (POS) ambiguity is a metric to quantify how frequent the dominant POS of a given word is given all possible POS a word has. The value is a fraction between 0 and 1 where 1 denotes that the word always occurs in its most dominant POS form. POS values were obtained from the SUBTLEXus database (Brysbaert et al., 2012).

Norms were available for 74,286 unique words.

Obtained from the SUBTLEXus database: https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus).

**Semantic Diversity** Semantic diversity is a metric that quantifies semantic ambiguity based on variability in the contextual usage of words. The metric was computed by (Hoffman et al., 2013) and takes a step beyond simply summing the number of definitions that a word has. The underlying assumption is that words that appear in a wide range of contexts on diverse topics are more variable in meaning than those that appear in a restricted set of similar contexts. Hoffman et al. thus quantify the degree to which the different contexts associated with a given word vary in their meanings.

Norms were available for 31,739 English words. Obtained from: https://link.springer.com/article/10.3758/s13428-012-0278-x#SecESM1.

**Word Length** Word length as measured by characters.

**Zipf Lexical Frequency** The Zipf lexical frequency is a metric of word frequency, but on a different scale than standard frequency. The Zipf scale was proposed by (van Heuven et al., 2014) as a scale that is easier to interpret than the usual frequency scales. Zipf values range from 1 to 7, with the values 1-3 indicating low-frequency words (with frequencies of 1 per million words and lower) and the values 4-7 indicating high-frequency words (with frequencies of 10 per million words and higher). Norms were based on American subtitles (51 million words in total) from the SUBTLEXus database (Brysbaert and New, 2009).

Norms were available for 74,286 unique words. Obtained from the SUBTLEXus database: https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus).

### A.2 Contextual Features

The contextual features broadly fall into two categories: syntactic (denoted by Contextual_syntax) and miscellaneous (denoted by Contextual_misc). Contextual features are ordered alphabetically.

#### A.2.1 Contextual Features — Syntax

**Content Word Ratio — Misc** Lexical density is the proportion of content words to function words in a sentence. It is a proxy for how much information a sentence contains. Content words were defined as nouns, verbs, adjectives, and adverbs and were defined using the NLTK part-of-speech tagger.

**Dependency Locality Theory (DLT) Variants** The Dependency Locality Theory (DLT) (Gibson, 2000) features are measures of storage and integration costs during sentence processing. The DLT is a theory of word-by-word comprehension difficulty during human language processing, with difficulty hypothesized to arise from working memory demand related to storing items in working memory (storage cost) and retrieving items from working memory (integration cost) as required by the dependency structure of the sentence. We include the traditional DLT metrics, as well as modifications as described in (Shain et al., 2016).

**Left-Corner Features — Syntax** The left-corner features are based on left-corner theories of sentence processing as described in (Rasmussen and Schuler, 2018). Similar to DLT, left-corner parsing models also posit storage and integration costs, but these costs are thought to derive not from dependency locality but from the number of unconnected fragments of phrase structure trees that must be maintained and combined in memory throughout parsing, word-by-word. See (Shain et al., 2021a) for detailed description of these features, but in brief they include: end of constituent, length of constituent (3 variants), end of center embedding, start of multi-word center-embedding, end of multi-word center embedding, length of multi-word center embedding (3 variants), and syntactic embedding depth. Features are derived from automatic parse trees generated by the van Schijndel et al. (2013) parser trained on a generalized categorial grammar reannotation (Nguyen et al., 2012) of the Penn Treebank corpus (Marcus et al., 1993).

**N-gram Surprisal — Misc** N-gram surprisal provides a metric of how surprising a word is given its context. The norms were computed by (Piantadosi et al., 2011) based on Google (Brants and Franz, 2009) using a standard probabilistic N-gram model which treats the context as consisting only of the local linguistic context containing the previous $N-1$ words. The norms are available for $N = 2, 3, 4$, i.e. 2-grams, 3-grams and 4-grams.

Norms were available for 3,297,629 (2-grams), 2,133,709 (3-grams) and 1,600,987 (4-grams) unique words. Obtained from colala.berkeley.edu/data/PiantadosiTilyGibson2011/Google10L-1T.

**Pronoun Ratio — Misc** The pronoun ratio is

the proportion of pronoun words to all words in a sentence. It is a proxy for how much discourse is assumed in a sentence. Pronoun words were defined using the NLTK part-of-speech tagger.

The following features were not analyzed in the current work, but are in the process of being added to the `SentSpace` framework:

**Language Model Surprisal** Language model surprisal provides a metric for how surprising (i.e., likely) a given sentence is by using the probability distribution obtained from pre-trained state-of-the-art language models. The default probability is computed as the product of individual tokens' log probabilities.

The language models were obtained using the HuggingFace Transformers framework (Wolf et al., 2020).

**Sentence-Level Sentiment — Misc** Sentence-level sentiment provides a metric for how positive or negative a given sentence is. The feature was derived using a pre-trained transformer model fine-tuned to perform sentiment prediction from a large dataset of human-annotated sentiment norms. The code framework used to compute the feature was by HuggingFace (Wolf et al., 2020).

**Syntactic Rule Frequency — Syntax** The syntactic rule frequencies consist of counts of n-ary and binary syntactic rules. For both n-ary (n is an arbitrary number larger than two) and binary rules, the sentence is dependency parsed (CoNLL format). The n-ary version gets all heads, along with its part of speech and its dependency relations. So if there is a verb with a subject and object, an n-ary rule would be: nsubj-V-dobj. The binary version is similar, but takes each head/dependency relationship separately. So the rules would be nsubj-V, V-dobj, etc. Thus, the syntactic rule frequency features consists of the number of times each $n$-ary and binary rules occur. The code framework to compute the features was obtained from Kyle Mahowald (*preprint to appear*).

### A.3 Embeddings

Embeddings provide representations of sentences in high-dimensional, learned vector spaces. We provide a context-independent embedding space (words have the same vector representation independent of context), GloVe (Pennington et al., 2014), as well as several commonly used contextualized embedding spaces (words have different vector representations based on the context in which they appear) from the HuggingFace framework (Wolf et al., 2020). These can be invoked using the corresponding model identifier from HuggingFace model hub as an argument to `SentSpace`.

## B Materials Used in Demonstration

### B.1 Example human-written paragraph

Montreal has a unique bagel tradition that dates back to at least 1919. The Montreal bagel is chewier, smaller, and less dense, but has a much bigger hole than its American cousin. Boiled in honey water and then baked in a wood-burning oven, it's a little sweet and has a harder exterior. It is hand-rolled in the shape of an oval hoop; you can wear one around your wrist like a bracelet.

### B.2 Examples of GPT2-XL generated text

**1** Montreal has a unique bagel tradition that dates back to the mid-nineteenth century. In the early days of the morning, tourists would arrive in Montreal in large numbers for a morning sandwich and coffee. You could be guaranteed to find several dozen different varieties in a pastry case. When the "Cappuccino" was introduced in 1894, the bagels came with it. You had to order your espresso and and your bagel separately.

**2** Montreal has a unique bagel tradition that dates back to 1766, when an Englishman named J.D. Bagel opened the first of what would become five bakeries. Today, there are more than 70 Bagel East St. Laurent and 15 Bagel Central St. Laurent. It's not just the flavors of bagels that separate Montreal bagel restaurants from the rest of Canada. What also sets them apart are the quality ingredients.

**3** Montreal has a unique bagel tradition that dates back to 1876, when the city hosted the Olympic Games. The tradition is still alive today. Because the bagel is such an American food, you won't find bagels being served in many Chinese restaurants, as French food in France is considered food that originated from the Americas. Desserts Although bagels have many tastes, you don't need to try

every flavor to enjoy bagel ice cream or bagel sandwiches.

**4** Montreal has a unique bagel tradition that dates back to 1917. In 2017, the Bagel Corner will open its doors again to celebrate Montreal's unique bagel heritage. This May, the Bagel Corner will offer a variety of bagels, bagel sandwiches and doughnuts, sold in Montreal and Montreal-area food trucks. With bags of fresh fruit, grapes, pickled veggies and other delicious bagels, this year's sale will be a great opportunity to indulge your taste buds.

**5** Montreal has a unique bagel tradition that dates back to the early 1900's. The bagels that we now associate with Montreal are greatly influenced by the place that spawned them the Sea Route Bagel. This famous bagel began its long journey to Montreal with a group of Jewish immigrants arriving from Eastern Europe in the early 1900's. To say that they were fortunate would be an understatement.

## C Corpora Benchmarks

As mentioned in the manuscript, we subsample a collection of corpora for use as benchmarks to compare against. We list these corpora in Table 1. These include: the Brown Corpus (Francis and Kucera, 1979), the Toronto Books Corpus Adventure genre (Zhu et al., 2015), Wall Street Journal corpus (Paul and Baker, 1992), Universal Dependencies https://universaldependencies.org/#download, Colossal Cleaned Common Crawl (C4) (Raffel et al., 2020), Corpus of Contemporary American English (Spoken) — 1991, 2001, 2012 (Davies, 2009).

| Corpus name | Volume/Genre | No. of sentences | Duplicates | Sentence Length | | |
|---|---|---|---|---|---|---|
| | | | | Mean | Median | Std. deviation |
| (Total) | | 63,350,596 | 615,672 | | | |
| Brown | | 26,954 | 3 | 12.681 | 13 | 4.462 |
| Toronto Books | Adventure | 1,040,936 | 0 | 14.519 | 13 | 7.632 |
| UD | | 19,543 | 2 | 16.89 | 15 | 8.664 |
| WSJ | | 541,790 | 960 | 22.351 | 21 | 8.734 |
| COCA Spoken | 1991 | 121,351 | 0 | 17.051 | 15 | 10.048 |
| COCA Spoken | 2001 | 113,330 | 3 | 15.734 | 13 | 9.269 |
| COCA Spoken | 2012 | 97,512 | 2 | 14.41 | 12 | 8.65 |
| C4 | 10 random parts | 49,772,404 | 614,602 | 17.481 | 16 | 8.905 |

Table 1: A list of Corpora used as Benchmarks

| Feature | Mean | | Standard Deviation | |
|---|---|---|---|---|
| | GPT2-XL | Human | GPT2-XL | Human |
| 2-gram Surprisal*** | 7.86 | 8.13 | 0.46 | 0.40 |
| 3-gram Surprisal*** | 5.64 | 5.78 | 0.32 | 0.28 |
| 4-gram Surprisal | 3.57 | 3.58 | 0.22 | 0.20 |
| Age of Acquisition | 5.28 | 5.26 | 0.53 | 0.46 |
| Arousal | 0.42 | 0.42 | 0.04 | 0.04 |
| Concreteness*** | 2.58 | 2.69 | 0.16 | 0.15 |
| Content Word Ratio* | 0.50 | 0.53 | 0.11 | 0.08 |
| End of Constituent* | 0.23 | 0.24 | 0.03 | 0.03 |
| End of Center-Embedding | 0.65 | 0.64 | 0.04 | 0.04 |
| End of Multi-Word Center-embedding | 0.12 | 0.13 | 0.03 | 0.02 |
| Length of Constituent* | 1.33 | 1.38 | 0.18 | 0.12 |
| Length of Multi-Word Center-Embedding | 0.47 | 0.52 | 0.17 | 0.13 |
| Lexical Connectivity | 45.12 | 43.85 | 7.54 | 7.48 |
| Lexical Decision RT | 626.59 | 629.34 | 11.02 | 9.30 |
| Log Contextual Diversity* | 3.46 | 3.40 | 0.17 | 0.14 |
| Log Lexical Frequency*** | 4.48 | 4.36 | 0.27 | 0.20 |
| Number of Morphemes | 1.52 | 1.54 | 0.09 | 0.08 |
| Orthographic Neighbor Frequency | 690.82 | 655.06 | 228.56 | 196.97 |
| Orthography-Semantics Consistency | 0.77 | 0.76 | 0.04 | 0.04 |
| Polysemy | 5.49 | 5.46 | 0.86 | 0.69 |
| Prevalence | 2.30 | 2.31 | 0.04 | 0.03 |
| Pronoun Ratio | 0.07 | 0.07 | 0.07 | 0.07 |
| Sentence Length | 104.59 | 102.63 | 12.50 | 9.59 |
| Syntactic Integration Cost* | 0.62 | 0.68 | 0.17 | 0.14 |
| Syntactic Embedding Depth | 1.23 | 1.27 | 0.17 | 0.14 |
| Valence | 0.62 | 0.62 | 0.05 | 0.04 |

Table 2: Mean values of `SentSpace` features for human- and GPT2-XL-generated text. Statistically significant differences (after a two-tailed $t$-test) are indicated by $*p < .05$, $**p < .01$, $***p < .005$.