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

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

What is SentSpace?

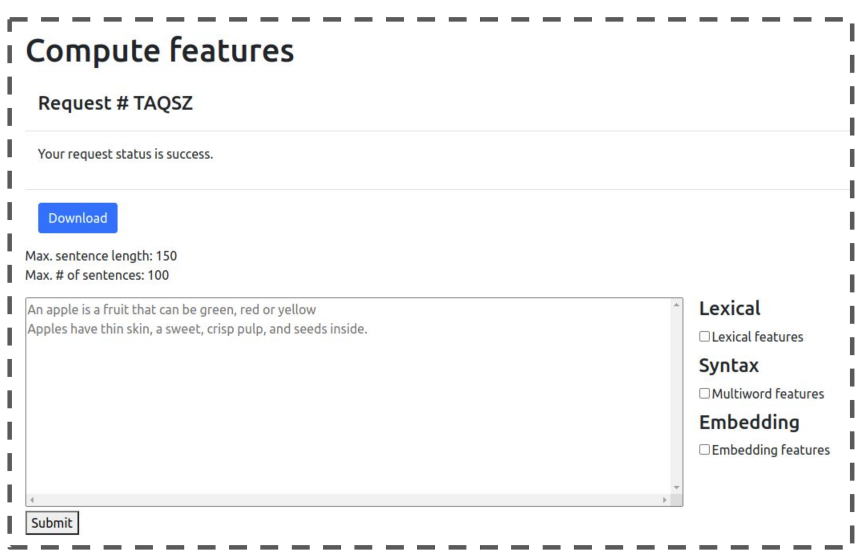
- SentSpace is a modular, open-source framework for streamlined evaluation of text.
- SentSpace characterizes textual input using cognitively motivated lexical, syntactic, and semantic features.
- Features are derived from psycholinguistic experiments, large-scale corpora, and theoretical proposals.
- Core sentence features fall into two primary feature spaces:
  1) *Lexical*
  2) *Contextual/Syntactic*
- SentSpace can be accessed from a web interface or a Python package.
- 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.
- SentSpace provides a broad set of cognitively motivated linguistic features for evaluation of text within natural language processing, cognitive science, and the social sciences.

Usage

1. Command-Line Interface (CLI)

```
python -m sentspace
  input -lex 1 -syn 1
  -usermodule 1 -o
output.csv
```

2. Hosted Frontend



sentspace.github.io/hosted

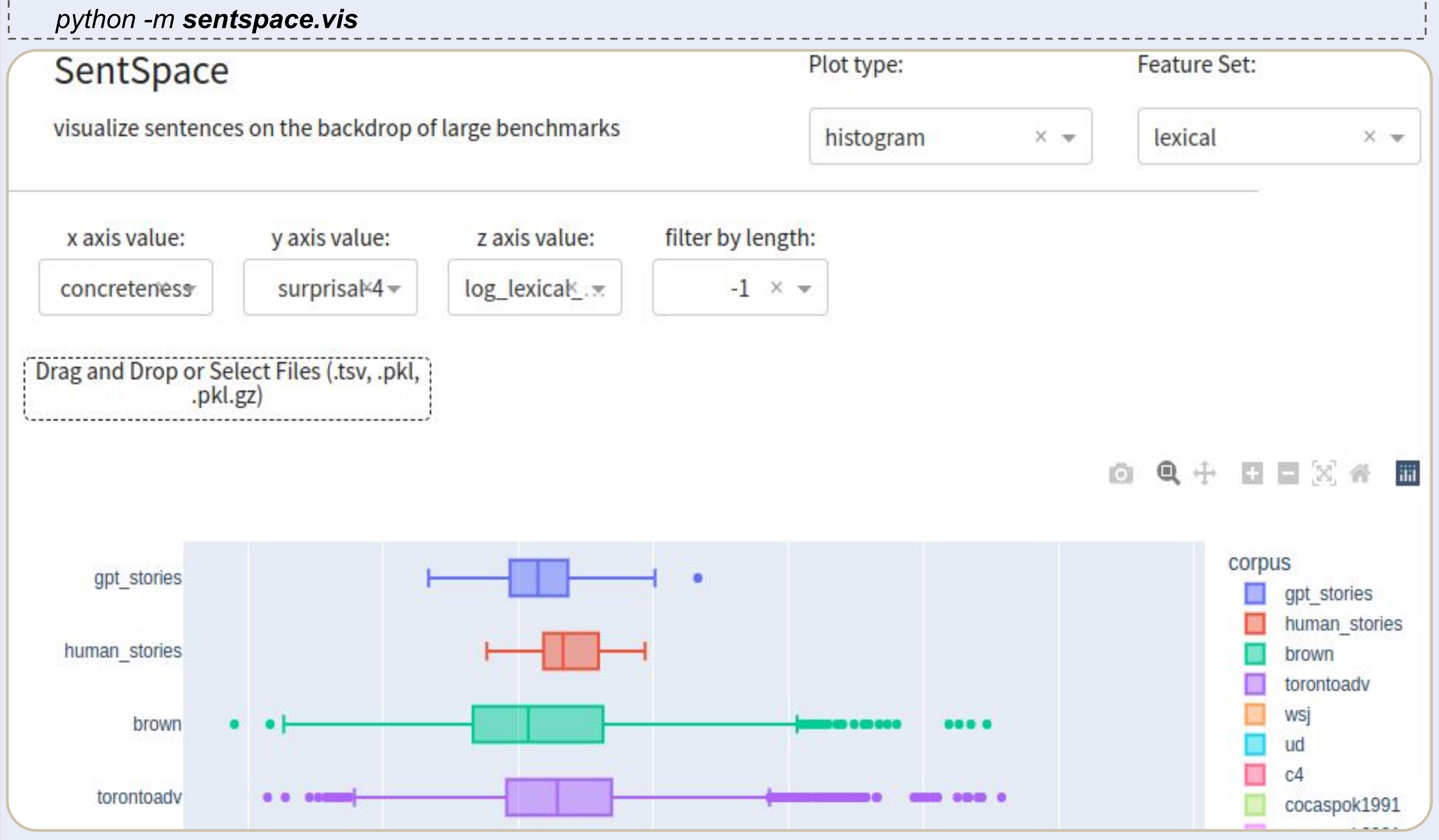
hosted web app

sentspace.github.io/sentspace

API documentation

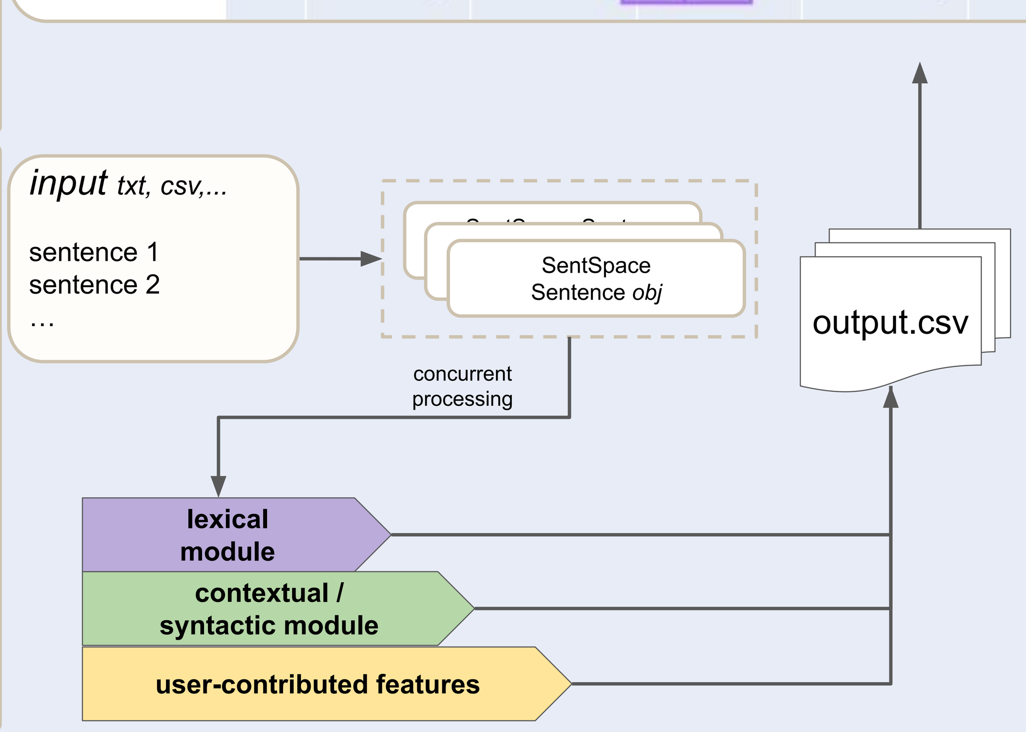
Visualization

```
python -m sentspace.vis
```



Structure

Data Flow



SentSpace Features

$f(\text{sentence}) \mapsto \mathbb{R}^n$

At its core, SentSpace organizes features into two main modules based: *Lexical* & *Contextual/Syntactic*

lexical module

- Age of Acquisition (Kuperman et al., 2012)
- Arousal (Mohammad, 2018)
- Body-Object Interaction (Pexman et al., 2019)
- Concreteness (Brysbaert et al., 2014)
- Contextual Diversity (SUBTLEXus: Brysbaert & New, 2009)
- Dominance (Mohammad, 2018)
- Imageability (Scott et al, 2019)
- Lexical Connectivity (Mak & Twitchell, 2020)
- Lexical Decision Latency (Balota et al., 2007)
- Lexical Frequency (SUBTLEXus: Brysbaert & New, 2009)
- Number of Morphemes (Morfessor: Virpioja et al., 2013)
- Orthographic Neighbor Frequency (Medler & Binder, 2005)
- Orthographic-Semantics Consistency (Marelli & Amenta, 2018)
- Polysemy (Miller, 1992)
- Prevalence (Brysbaert al., 2019)
- Sensorimotor norms (11 different norms) (Lynott et al., 2020)
- Socialness (Diveica et al., 2022)
- Valence (Mohammad, 2018)

contextual / syntactic module

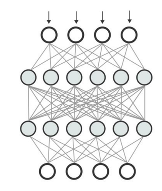
- Dependency Locality Theory (DLT) (Gibson, 2000):
  - Various features that quantify storage and integration cost based on the dependency structure of the sentence.
- Left-corner features (Rasmussen & Schuler, 2018):
  - Various features derived from a left-corner parser such as center embedding depth and constituent lengths.
- N-gram surprisal (Piantadosi et al., 2011)
- Part of Speech ratios
  - Content word ratio, pronoun ratio

Comparison Between Machine- and Human-Generated Text

Open source experiment code:
https://github.com/sentspace/NAACL-HLT-2022

- **Question:** Can we reveal quantitative differences between GPT2-XL-generated and human-generated text?
- **Approach:** Generate text using artificial language models (GPT2-XL) and humans:

Prompt:
"*Montreal has a unique bagel tradition that dates back to*"



GPT2-XL

```
… 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.
```

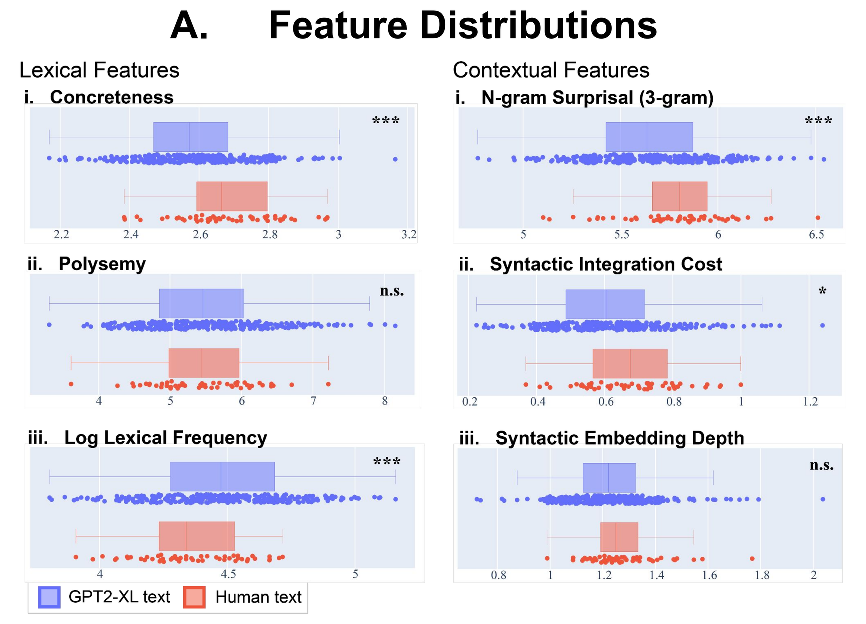
```
… at least 1919. The Montreal bagel is chewier, smaller, and less dense, but has a much bigger hole than its American cousin. Boiled in honey water and then baked in a wood-burning oven, it's a little sweet and has a harder exterior. It is hand-rolled in the shape of an oval hoop; you can wear one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

Human

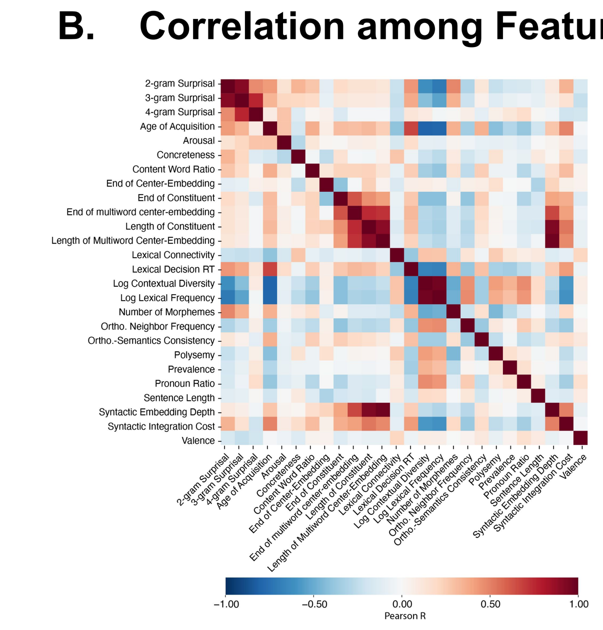
52 unique 10-word prompts (GPT2-XL: 5 paragraphs per prompt; Human: 1 paragraph per prompt)

Obtain SentSpace sentence-level features and compare GPT2-XL and humans

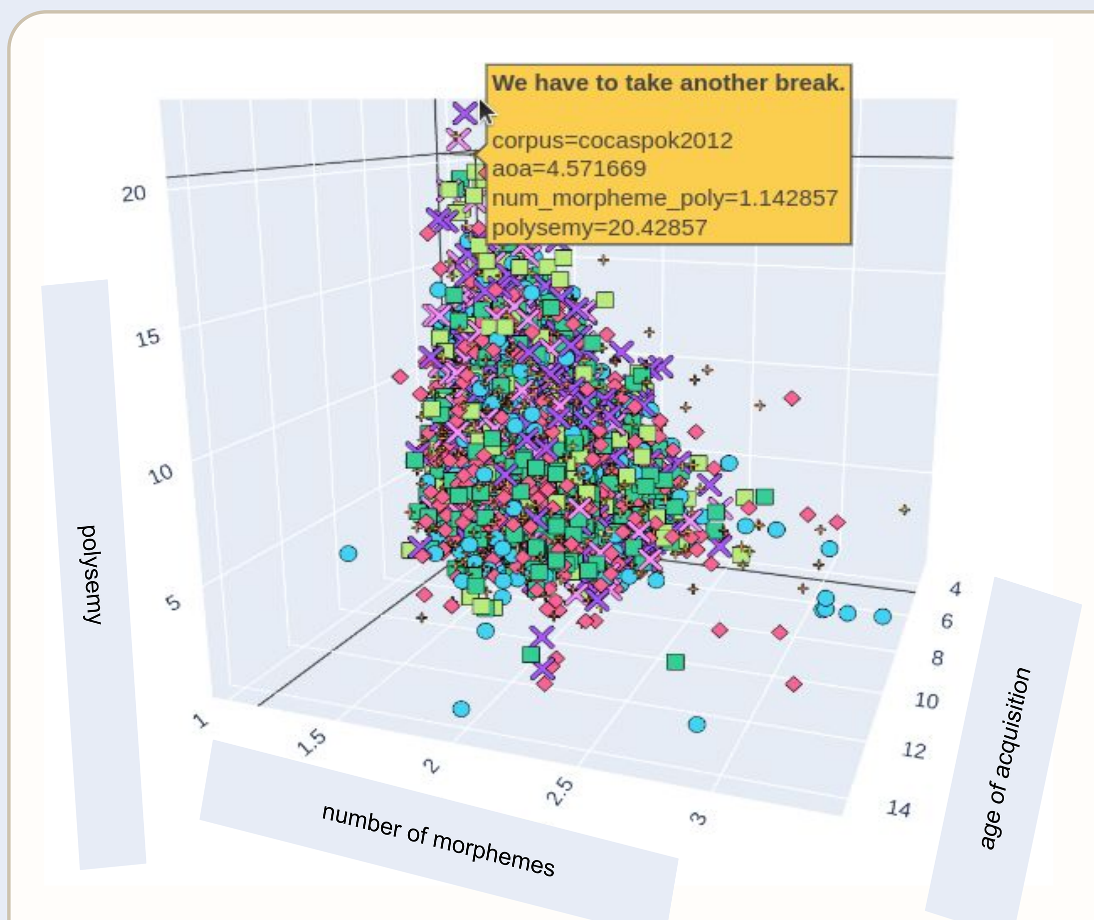
A. Feature Distributions



B. Correlation among Features



- **Conclusion:** GPT2-XL-generated text appears fluent at the surface level, but our features can reveal subtle differences between GPT2-XL and human-generated text: For instance, GPT2-XL produced less concrete sentences with shorter syntactic dependencies.



Other Use Cases

- Evaluation of text used for training of language models
- Probing high-dimensional representations from ANNs
- Comparison of text produced by different human populations (e.g., neurotypical and individuals with communication disorders or kids)
- Comparison of different genres of text
- Analysis of fine-grained variation in human behavioral and neural responses with respect to sentence features

Extending SentSpace

user-contributed features

- Simple token-level feature addition from a CSV

```
python -m sentspace.package_lexical
  input.csv  --word_column Word   --feature_column LDRT
    --feature_name lexical_decision_latency
```

- Features requiring computation:
  - Create a module following SentSpace API
  - `Callable[sentspace.Sentence,Dict]`

```
→ {
      index: …, token: …,
      sentence: …, feature_name: val,
      …
  }
```

PRs welcome

Acknowledgements & References

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. G.T. is grateful for funding from the International Doctoral Fellowship from AAUW. We also thank an R01 award DC016607 from NIDCD and a U01 award NS121471 from NINDS.

Brysbaert et al. (2014): *Beh Res Methods*, 46(3):904–911. **Brysbaert et al. (2019)**: *Beh Res Methods*, 51(2):467–479. **Brysbaert & New (2009)**: *Beh Res Methods*, 41(4):977–990. **Diveica et al. (2022)**: *Beh Res Methods*, 1-13. **Gibson (2000)**: *Image, Language, Brain*, 94-126. **Kuperman et al. (2012)**: *Beh Res Methods* 44(4):978-90. **Mak & Twitchell (2020)**: *Psych Bull Rev*, 27(5):1059–1069. **Marelli & Amenta (2018)**: *Beh Res Methods*, 50(4):1482–1495. **Medler & Binder (2005)**: *neuro.mcw.edu/mcword*. **Miller (1992)**: *Comm. ACM*, 38:39–41. **Mohammad (2018)**: *ACL 2018*. **Pexman et al. (2019)**: *Beh Res Methods*, 51(2), 453-466. **Piantadosi et al. (2011)**: *PNAS*, 108(9):3526–3529. **Rasmussen & Schuler (2018)**: *Cog Sci*, 42 Suppl 4:1009-1042. **Virpioja et al. (2013)**: *Aalto University publication*, 978-952-60-5501-5.