Data in Brief 46 (2023) 108799

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Semantic Coherence Dataset: Speech transcripts

Davide Colla, Matteo Delsanto, Daniele P. Radicioni*

University of Turin, Italy

ARTICLE INFO

Article history:
Received 4 October 2022
Revised 21 November 2022
Accepted 28 November 2022
Available online 2 December 2022

Keywords:
Perplexity metrics
Intra-subject semantic reliability
Inter-subject semantic reliability
Language models
Speech transcripts
Spoken language analysis

ABSTRACT

The Semantic Coherence Dataset has been designed to experiment with semantic coherence metrics. More specifically, the dataset has been built to the ends of testing whether probabilistic measures, such as perplexity, provide stable scores to analyze spoken language. Perplexity, which was originally conceived as an information-theoretic measure to assess the probabilistic inference properties of language models, has recently been proven to be an appropriate tool to categorize speech transcripts based on semantic coherence accounts. More specifically, perplexity has been successfully employed to discriminate subjects suffering from Alzheimer Disease and healthy controls. Collected data include speech transcripts, intended to investigate semantic coherence at different levels: data are thus arranged into two classes, to investigate *intra-subject* semantic coherence, and *inter-subject* semantic coherence. In the former case transcripts from a single speaker can be employed to train and test language models and to explore whether the perplexity metric provides stable scores in assessing talks from that speaker, while allowing to distinguish between two different forms of speech, political rallies and interviews. In the latter case, models can be trained by employing transcripts from a given speaker, and then used to measure how stable the perplexity metric is when computed using the model from that user and transcripts from different users. Transcripts were extracted from talks lasting almost 13 hours (overall

DOI of original article: 10.1016/j.artmed.2022.102393
* Corresponding author.
E-mail address: daniele.radicioni@unito.it (D.P. Radicioni).
Social media: @DavideColla6 (D. Colla), @d_radicioni (D.P. Radicioni)

https://doi.org/10.1016/j.dib.2022.108799
2352-3409/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

12:45:17 and 120,326 tokens) for the former class; and almost 30 hours (29:47:34 and 252,270 tokens) for the latter one. Data herein can be reused to perform analyses on measures built on top of language models, and more in general on measures that are aimed at exploring the linguistic features of text documents.

© 2022 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Specifications Table

Subject	Artificial Intelligence.
Specific subject area	Collected data are for experiments in Natural Language Processing (NLP), to test metrics conceived to assess the semantic coherence of text documents.
Type of data	Text Code
How the data were acquired	Transcripts of spoken language were downloaded from different websites. A script was devised to download transcripts of speeches from eight well-known past and present political figures: Joe Biden, Bill Gates, Boris Johnson, Martin Luther King, Nelson Mandela, Barack Obama, Bernie Sanders, Donald Trump. Data were arranged into two classes, that are intended for testing intra-subject semantic coherence (10 transcripts of talks from a single speaker), and inter-subject semantic coherence (40 transcripts from 8 different speakers).
Data format	Raw Code
Description of data collection	Transcripts in the former class (intended for analyzing spoken language in an *intra-subject* perspective) were kept as homogeneous as possible and include speeches from interviews and campaign rallies; data in the latter class (used to compare language models acquired from different speakers, and thus employed in *inter-subject* experiments) are more varied and contain transcripts of speeches on spot topics, such as economy, health systems, civil rights and so forth.
Data source location	The transcripts were collected from the following websites: http://db.nelsonmandela.org https://blogs.lse.ac.uk https://news.harvard.edu https://prorhetoric.com https://www.c-span.org https://www.conservatives.com https://www.crmvet.org https://www.gatesfoundation.org https://www.gov.uk https://www.rev.com https://www.sbs.com.au https://www.smu.edu https://www.vox.com https://www.weforum.org https://www.whitehouse.gov
Data accessibility	Mendeley data: *Semantic Coherence Dataset - SCD* [2] Data identification number: 10.17632/s4dtmfmzxw.1 Direct URL to data: https://data.mendeley.com/datasets/s4dtmfmzxw/1 Zenodo data: *Semantic Coherence Markers: Code* [3] Data identification number: 10.5281/zenodo.7118402 Direct URL to data: https://zenodo.org/record/7118402#.YzQnxS8RqgR
Related research article	Colla, D., Delsanto, M., Agosto, M., Vitiello, B., & Radicioni, D. P. (2022). Semantic coherence markers: The contribution of perplexity metrics. *Artificial Intelligence in Medicine*, 134, 102393.

Value of the Data

- To date, language models are being used to solve many NLP tasks. Providing evidence that the metrics employed to deal with semantic coherence of text documents are reliable is therefore a prerequisite for many tasks. These data were used to assess the perplexity metrics, which was then employed to discriminate between the transcripts of speeches from healthy subjects and subjects suffering from Alzheimer Disease [1,2].
- The questions one may be able to answer by training and testing language models on these data are: are we able to reliably detect the difference in linguistic register between interviews and political rallies? Are we able to assess the coherence of a given speaker with her/his past speeches? Are we able to assess the coherence of the speeches of the eight selected speakers (whose transcripts were compiled in this dataset)? Researchers will be able to fine-tune and test their models on these data to answer these and further questions.
- Although these data were originally collected for experimenting with the perplexity metrics (a measure originally conceived to assess language models), different approaches can be envisaged to measure the semantic coherence of speeches herein, and their reliability may be compared with that computed through the perplexity.
- The released data along with the implemented system (also publicly available [3]) offer a testbed and a baseline for measuring the reliability of statistical measures of semantic coherence.

1. Objective

The data in this dataset can be used to investigate the reliability of the perplexity metrics and of language models [1]. Given a word sequence of κ elements, $W = \{w_1,\ w_2,\ ...,\ w_k\}$, and a language model LM, the perplexity of LM and W is defined as

$$PPL(LM, W) = exp\left\{-\frac{1}{k}\sum_{i=1}^{k}\log LM(w_i \mid w_{1:i-1})\right\}.$$

We thus can see that low PPL values indicate that the model is able to closely predict the sequence W; on the contrary, high PPL scores, corresponding to low probability values, indicate that the model is 'perplexed' and unable to predict W (more on perplexity in [1]).

Collected data allow recording two different senses of coherence. *Intra-subject* coherence, featuring the speeches from a given speaker, can be measured by acquiring different models for specific types of talk (e.g., interviews and political rallies) and then comparing perplexity scores obtained in same condition *vs.* cross-condition. Likewise, in the *across-subject* coherence we may investigate whether the model acquired from talks of a given speaker is compatible (i.e., featured by low PPL scores) with the utterances from the other speakers.

2. Data Description

Data are arranged into two directories, for measuring the intra-subject and inter-subject reliability of the perplexity metrics, respectively. Transcripts collected have been extracted from almost 13 hours of talks (overall 12:45:17; 10 talks from a single speaker; 120,326 tokens) for the former class, and from almost 30 hours of talks (29:47:34; 40 talks from 8 different speakers; 252,270 tokens) for the latter one.

The data in the first directory contain ten transcripts from talks of the former US President Donald Trump, five interviews and five campaign rallies, all recorded between June 2019 and November 2020. The statistics describing all transcripts employed in the first experimental setting, including time duration, token counts and type-token ratio (TTR, computed as the ratio

Table 1
Statistics describing the transcripts employed in Experiment 1: for all considered samples we report time duration, number of tokens, number of unique tokens, average number of tokens and of unique tokens, and type-token ratio (TTR).

Category	Transcript	Duration	Tokens	Unique Tokens	AVG Tokens	AVG Unique Tokens	TTR
Interview	I	01:28:52	7278	1098	8953	1185	0.13
	II	01:28:23	6471	922			
	III	01:31:34	18,514	1926			
	IV	00:45:40	6702	1032			
	V	01:01:51	5933	946			
Rally	I	01:17:37	15,200	1967	15,051	1944	0.13
	II	00:56:17	10,501	1614			
	III	01:43:43	20,865	2300			
	IV	01:13:01	14,056	1945			
	V	01:18:19	14,806	1896			

between the types, that is the total number of different tokens occurring in a text divided by the total number of tokens) are reported in Table 1. Two different kinds of discourse were targeted here: the interview and the political rally. While in the former case both the questions put to the interviewee and the answers may cover different topics, political rallies are events where people sharing similar political beliefs gather to support their candidate. In this case the language adopted is in principle more regular, and not concerned with answering specific questions. As regards the linguistic register differentiating such transcripts, interviews should convey a sense of poise, balance and posture, while the language adopted in rallies is expected to be more emphatic, direct, uniform and vehement. Language models fine-tuned on such data should grasp such differences [1].

The data in the latter directory were employed to investigate whether perplexity scores are stable across subjects, by measuring to what extent the LMs acquired from talks from a speaker fit to the talks of different speakers. We have collected transcripts of speeches from eight well-known past and present political figures: Joe Biden, Donald Trump, Barack Obama, Bernie Sanders, Bill Gates, Nelson Mandela, Martin Luther King, and Boris Johnson. The statistics describing each transcript herein also include time duration, token counts and type-token ratio are reported in Table 2.

The source code employed in the experimentation of [1] is publicly available on Zenodo [3]. The library allows running the two experiments on the Semantic Coherence Dataset.

More specifically, the code developed for the first experiment (*Intra-subject and discourse-level coherence*) is to compute perplexity scores for a given pair (LM, input text) to test whether they are stable, and whether perplexity scores are able to grasp factors specific to a given sort of speech. To investigate reliability, we recorded the *coefficient of variation (CV)* metric, which is computed as the ratio between standard deviation of the perplexity scores and the average of perplexity scores, under the assumption that low CV scores (CV $\leq$ .1) support the hypothesis that perplexity provides stable and reliable scores. The testing facilities implement a one-speech-out approach, which is described in detail in [1].

The code developed for the second experiment (*Inter-subject coherence on different speakers*) is to compute perplexity scores to test whether such scores are stable across subjects. The experimental setting implemented herein considers five transcripts for eight well-known past and present political figures and acquires a language model for each subject. The perplexity scores for the speeches from each speaker are then computed based on all others' language models. To investigate this sort of reliability, we recorded Intraclass correlation coefficients (ICC) [4], under the assumption that ICC values above 0.9 indicate excellent reliability [5].

All deployed source code contains packages to run the experiments with both GPT-2 and N-grams, whose smoothing employs the interpolated Kneser–Ney Smoothing technique [6].

Table 2
Figures describing the transcripts employed in Experiment 2: time duration, number of tokens, number of unique tokens (along with average number of tokens and average number of unique tokens) and type-token ratio (TTR) are reported for each such speech transcript.

Subject	Transcript	Duration	Tokens	Unique Tokens	AVG Tokens	AVG Unique Tokens	TTR
Joe Biden	I	0:32:23	4647	1074	6315	1343	0.21
	II	0:41:39	5446	1140			
	III	0:25:00	9490	1895			
	IV	0:43:36	6801	1381			
	V	0:34:05	5211	1226			
Donald Trump	I	1:17:37	15,200	1967	15,051	1185	0.13
	II	0:56:17	10,501	1614			
	III	1:43:43	20,865	2300			
	IV	1:13:01	14,056	1945			
	V	1:18:19	14,806	1896			
Barack Obama	I	0:56:39	5594	1479	5957	1271	0.21
	II	0:38:15	6298	1252			
	III	0:38:45	5526	1153			
	IV	0:45:55	6981	1312			
	V	0:36:07	5390	1159			
Bernie Sanders	I	0:35:33	4164	969	4458	1046	0.23
	II	0:29:51	3785	849			
	III	0:34:54	4451	1088			
	IV	0:43:27	5387	1039			
	V	0:44:46	4501	1286			
Bill Gates	I	0:35:53	3503	944	2514	812	0.32
	II	0:17:20	1679	577			
	III	0:24:07	2350	779			
	IV	0:22:04	2152	744			
	V	0:30:07	2896	1018			
Nelson Mandela	I	0:40:17	3844	1113	6403	1410	0.22
	II	0:29:45	1740	617			
	III	3:00:00	15,682	2702			
	IV	1:43:21	7741	1654			
	V	0:40:16	3020	963			
Martin Luther King	I	0:42:51	5197	1102	6508	1379	0.21
	II	0:46:56	6471	1315			
	III	0:43:48	6287	1456			
	IV	0:40:38	8256	1697			
	V	0:47:54	6332	1324			
Boris Johnson	I	0:51:42	4397	1123	3202	943	0.29
	II	0:20:35	2758	764			
	III	0:17:47	1960	659			
	IV	0:17:00	2375	896			
	V	0:38:22	4530	1273			

3. Experimental Design, Materials, and Methods

Two scripts were implemented, one to download the mentioned material, and another one to extract the descriptive statistics presented in Tables 1 and 2, such as the number of tokens, unique tokens, the average number of tokens, the average number of unique tokens, and the type-token ratio featuring each transcript.

Since the dataset is intended for experimenting on spoken language, all phenomena possibly occurring in spoken language (such as blends, false starts, reiterations, interjections and filled pauses possibly present in the transcripts) were retained, so as to be able to inform models (e.g., through fine-tuning) on specificities characterizing the texts at hand. Data were thus preserved

in raw format: no form of text normalization was undertaken, that is, neither cleaning, stemming, lemmatization nor tokenization or any kind of pre-processing were applied to collected data.

No Data Augmentation (DA) approach was adopted (such, e.g., paraphrasing), to avoid injecting any kind of noise, and to include only actually uttered sentences. DA is customarily performed through static word embeddings [7,8], back translation [9], text generation [10], contextualized word embeddings [11] etc. Such processing can be easily done with off-the-shelf software libraries (see, e.g., [12]) according to specific application needs by starting from our data.

Ethics Statements

This work involves data collected from different web platforms hosting transcriptions of public speeches; data redistribution policies for publicly available text documents were complied with.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Semantic Coherence Markers: Code (Original data) (Zenodo).
Semantic Coherence Dataset - SCD (Original data) (Mendeley Data).

CRediT Author Statement

Davide Colla: Conceptualization, Methodology, Data curation, Investigation, Writing – original draft; **Matteo Delsanto:** Conceptualization, Methodology, Data curation, Investigation, Writing – original draft; **Daniele P. Radicioni:** Conceptualization, Investigation, Writing – original draft, Supervision.

Acknowledgments

The first author was partly supported by a grant provided by the University of Turin (PAUSE project) in the frame of the Public Engagement 2021 funding initiative. We are grateful to all Companies and Organizations listed in Section 'Data source location' for kindly sharing carefully transcribed materials: we are indebted to the following companies and organizations:
https://www.rev.com,
https://www.whitehouse.gov,
http://db.nelsonmandela.org,
https://news.harvard.edu,
https://www.gatesfoundation.org,
https://www.gov.uk,
https://www.news24.com,
https://www.c-span.org,
https://prorhetoric.com,
https://www.conservatives.com,

https://www.smu.edu,
https://www.crmvet.org,
https://blogs.lse.ac.uk,
https://www.sbs.com.au,
https://www.weforum.org,
https://www.vox.com.

References

[1] D. Colla, M. Delsanto, M. Agosto, B. Vitiello, D.P. Radicioni, Semantic coherence markers: the contribution of perplexity metrics, Artif. Intell. Med. 134 (2022) 102393.
[2] D. Colla, M. Delsanto, D.P. Radicioni, Semantic coherence dataset - SCD, Mendeley Data, v1, 2022. https://data.mendeley.com/datasets/s4dtmfmzxw/1
[3] D. Colla: Semantic coherence markers: Code (sep 2022). doi:10.5281/zenodo.7118402. 7118402, 2022.
[4] P.E. Shrout, J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability, Psychol. Bull. 86 (2) (1979) 420.
[5] D. Liljequist, B. Elfving, R.K. Skavberg, Intraclass correlation–a discussion and demonstration of basic features, PLoS ONE 14 (7) (2019) e0219854.
[6] R. Kneser, H. Ney, Improved backing-off for m-gram language modeling, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, IEEE, 1995, pp. 181–184. Vol. 1.
[7] J. Pennington, R. Socher, C.D. Manning, GloVe: global vectors for word representation, in: Proceedings of the Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.
[8] D. Colla, E. Mensa, D.P. Radicioni, LESSLEX: linking multilingual Embeddings to SenSe representations of Lexical items, Comput. Linguist. 46 (2) (2020) 289–333 pages.
[9] D.R. Beddiar, M.S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Soc. Netw. Media 24 (2021) 100153.
[10] N. Malandrakis, M. Shen, A. Goyal, S. Gao, A. Sethi, A. Metallinou, A.A. AI, Controlled text generation for data augmentation, Intell. Artif. Agents (2019) 90 EMNLP-IJCNLP 2019.
[11] J. Devlin, M.W. Chang, K. Lee, & K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[12] B. Li, Y. Hou, W. Che, Data augmentation approaches in natural language processing: a survey, AI Open 2022 (3) (2022) 71–90.