Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to facilitate research on Hebrew as a second language, as part of a broader research endeavor on crosslingual language varieties, i.e., how the knowledge of one language affects a person's performance in another language.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by the computational linguistics group at the University of Haifa, as part of an international project on "Crosslingual Language Varieties: A Multifaceted Investigation" (PIs: Prof. Anke Lüdeling (Humboldt University), Prof. Anat Prior and Prof. Shuly Wintner (University of Haifa)). The data were obtained from the National Institute for Testing and Evaluation (NITE).

Who funded the creation of the dataset?

Deutsche Forschungsgemeinschaft (DFG), grant no. 398186468. Additional funding by the Data Science Research Center at the University of Haifa.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset is a collection of argumentative essays in Hebrew authored by prospective higher-education students in Israel. Authors were either native speakers of Hebrew or learners of Hebrew with one of the following native languages: Arabic, French, or Russian. The essays by Hebrew native speakers were written as part of the Psychometric Test, a general test required for admission by most higher education institutions in Israel. The essays by non-native speakers were collected as part of the *YAEL* test: a Hebrew proficiency test required for examinees who chose to sit the Psychometric Test in a language other than Hebrew. Both tests are administered by NITE.

How many instances are there in total (of each type, if appropriate)?

The dataset contains 1000 essays authored by native speakers of each of the aforementioned languages (Hebrew, Arabic, French, and Russian). 4000 essays in total.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The essays authored by native speakers of Hebrew are a subset of essays written as part of the Psychometric Test in the years 2012 and 2017. The essays authored by native speakers of the other three languages are a subset of essays written as part of the Yael Test during the years 2011-2020. The essays were extracted by NITE from the larger sets semi-randomly. The only constraints were that the grades of the YAEL essays (as determined by NITE) were above a certain level and that the distributions of essay grades were similar across the three native languages of the authors.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

All the essays in the dataset were tokenized. In addition, about a third of the essays authored by non-speakers of Hebrew (1013 out of 3000) underwent manual error correction and annotation by one of three annotators. The dataset includes both the processed texts (in csv format) and the raw texts in Microsoft Word format.

Is there a label or target associated with each instance? If so, please provide a description.

Each essay in the dataset is identified by a unique (random) identifier label. The labels are listed in a separate file that provides some metadata for each essay.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

The metadata does not specify the sex of some of the YAEL authors, there is one with missing age, and many with unavailable psychometric scores, family income and parental education information.

Also, not all essays were annotated.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No known relationships.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No recommended splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

None that we are aware of.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

The dataset contains essays that were authored by prospective university students who probably did not give explicit informed consent to using the data for research. All essays we obtained were sentence-scrambled so that the original texts are obscured. No personal identification information is included in the dataset.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

To the best of our knowledge, no.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. The dataset is composed of essays written by examinees as part of the admission process to higher education institutions in Israel.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Subpopulations can be defined based on age, gender and native language.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Very unlikely. It is possible that some authors discussed personal experiences in their essays which, in combination with the metadata (e.g., age, sex, L1), might give clues as to their identity.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Metadata for essays in the YAEL sub-corpus includes the author's: native language (an indicator of ethnicity and, to some extent, religion), sex, age, and family income.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was directly observable: it is a collection of raw texts obtained from the institute (NITE) that collected them from the authors.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The essays were collected manually: the texts were hand-written by examinees in the Psychometric and YAEL tests. Then, the hand-written texts were typed by NITE.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The essays were extracted from larger sets of essays semi-randomly. We obtained an equal number of essays for each L1, with a similar distribution of essay scores across the three L1s.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

NITE was in charge of data collection; we do not know what practices it used.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The essays were collected at the time of their creation.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

Yes. The dataset is composed of essays written by examinees as part of the admission process to higher education institutions in Israel.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The essays in the dataset were provided by the institute (NITE) that collected them directly from the authors.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

We do not know whether participants gave consent to collection of data by registering for the test.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

We do not know.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Preprocessing included conversion of original texts from Microsoft Word format to plain text format and was followed by tokenization of the texts.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes. The raw essays in Microsoft Word documents are included in the dataset.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Preprocessing was done with Doc2Text and CPA (https://chengafni.wordpress.com/cpa/download/).

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, we ran some classification experiments on the data.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes: https://github.com/HaifaCLG/HebrewEssayCorpus

What (other) tasks could the dataset be used for?

The dataset could be used in various classification tasks and statistical analyses to reveal possible correlations between personal variables (e.g., native language, gender) and linguistic properties of the texts.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks)? If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

As is always the case with datasets that differentiate sub-groups of individuals, future users may try to identify (primarily linguistic) features of sub-groups, e.g., typical language use by speakers of one of the L1s represented in the dataset. Future users may also compute correlations between native language and other meta-data available in the dataset. It is hard to imagine how such correlations may be abused, but it is not impossible.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Yes. The dataset should not be used to create algorithms that automatically predict the score of new essays, whether such prediction relies on the text itself or on the meta-data associated with it. We have no reason to believe that such predictions will be reliable.

In general, the dataset is intended for research purposes only, and should not be used for applications.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The corpus is available for research purposes upon request and subject to signing a license agreement form.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

A tarball over email or as a download link.

When will the dataset be distributed?

Immediately.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Distribution of the dataset is subject to a license agreement, available here:
https://github.com/HaifaCLG/HebrewEssayCorpus

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No. We have permission from NITE to distribute the data subject to the license agreement.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

None that we are aware of.

Maintenance

Who is supporting/hosting/maintaining the dataset?

Shuly Wintner, on behalf of the Computational Linguistics Group at the University of Haifa.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

shuly@cs.haifa.ac.il

Is there an erratum? If so, please provide a link or other access point.

Not yet; future updates will be available on https://github.com/HaifaCLG/HebrewEssayCorpus

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Probably not.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

No specific plans.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

No such mechanism planned.