The Hebrew Essay Corpus
Curation, Annotation, Evaluation, Uses

Chen Gafni, Anat Prior, Shuly Wintner, University of Haifa, Israel

The Hebrew Essay Corpus

- 3000 argumentative essays authored by native speakers of Arabic, Russian, and French
- 1000 argumentative essays by Hebrew native speakers
- The native essays were authored as part of the psychometric exam
- The non-native ones were written for the *YAEL* test
- Different settings: prompts, allotted time, required length, year of test
- Still, as comparable as possible

Non-native essays: Metadata

- Author's L1, gender, age, year of exam
- Parental education; family income
- Prompt
- Essay score
- The only available metadata for the native speaker essays is the essay score

Annotation

- 1000 non-native essays were annotated
- Annotators are native speakers, linguistics graduates
- 50 were annotated by two annotators for evaluation
- How to identify "errors", and the notion of "target hypothesis"
- Enabling automatic processing

Use cases: Three classification tasks

- Distinguishing between native and non-native authors
- Identifying author's L1
- Predicting proficiency score
- Analysis of the features that support the classification

Conclusion

- The corpus, the tools used for processing it, the annotation schema and the guidelines to the annotators are all available for research proposes
- We expect them to be valuable resources for any investigation of Hebrew as a second language
- In particular, for investigating transfer effects from Arabic, French, and Russian

DFG Deutsche Forschungsgemeinschaft