Comprehensive Complexity Analysis of Large-scale Learner Corpora with the Common Text Analysis Platform

Xiaobin Chen,
xiaobin.chen@uni-tuebingen.de
Detmar Meurers,
dm@sfs.uni-tuebingen.de

Tübingen University

October 5, 2017

Linguistic analysis of texts

(Automatic) Linguistic analysis has been widely used for:

- assessing text readability
- modeling processing difficulty of sentences
- analyzing/scoring student writings
- comparing language typologies and their historical development
- attributing authorship
- identifying native languages
- detecting plagiarism
- assessing answers to questions
- predicting diseases
- ...

Existing tools for text analysis

A number of tools have been released in the past few years. e.g.

- Syntactic and Lexical Complexity Analyzers (Lu, 2010)
- Cohmetrix (McNamara et al., 2014)
- Suite of Linguistic Analysis Tools (Crossley et al., 2016a,b), also `http://www.kristopherkyle.com/tools.html`
- Computerized Propositional Idea Density Rater (Brown et al., 2008, CPIDR).
- ETS's TextEvaluator `https://texteval-pilot.ets.org/TextEvaluator/`
- Pearson's Reading Maturity Metric
- Text Analysis, Crawling, and interpretation Tool (Dehghani et al., 2016, TACIT)

Problems with existing tools

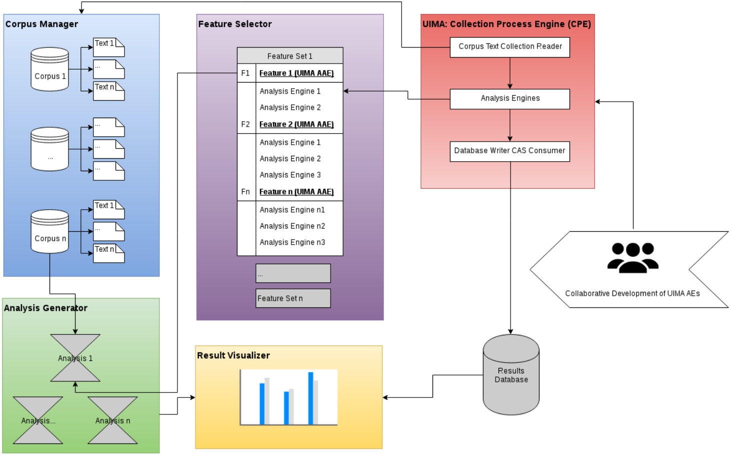
- Limited usability of tools and analysis components
  - OS-dependent standalone deployment
  - Source code release hard to use for non-programmers
  - Unfriendly user interface: command line interface, choice of features...
- Limited extensibility
  - Close source commercial systems
  - Non-reusable analysis components
- Collaborative development difficult to implement
  - Significant feature overlap
  - Duplication of efforts
- Feature proliferation, e.g.
  - CohMetrix: 106 metrics
  - Vajjala (2015): $>$200 features for readability assessment

System demands

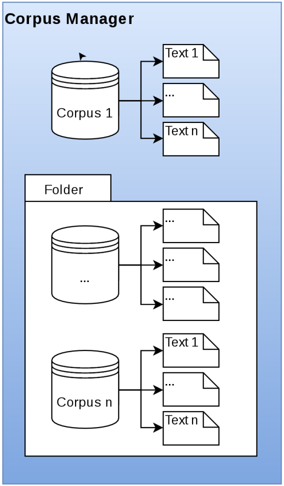
A system that is:

- Web-based
- user-friendly, supporting real-life use by ordinary users
- comprehensive set of linguistic features
- freedom to choose extracted features
- modularized and reusable analysis components

CTAP System Architecture



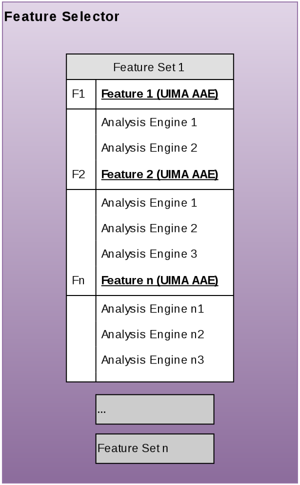
Corpus Manager



Helps users manage the language materials that need to be analyzed.

- Folders: grouping corpora
- Corpora: holding texts
- Tags: labeling texts based on e.g. document genre, target reader levels, etc.

Feature Selector

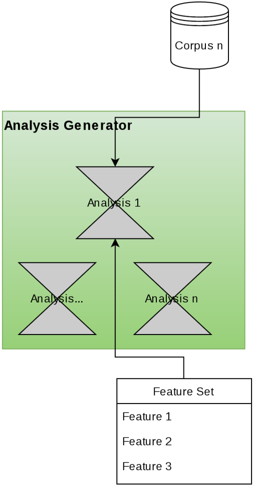


The Feature Selector supports:

- creating feature set to hold selected features
- add/remove features from feature set

Developers are encouraged to participate in in feature development at `https://github.com/ctapweb`.

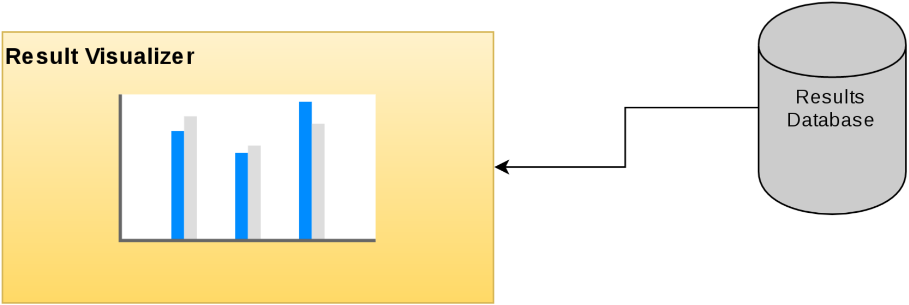
Analysis Generator



Each analysis extracts a set of features from the designated corpus. The analysis generator is used to:

- create new analyses
- run analyses and monitor their progress
- export analysis results in CSV format

Result Visualizer

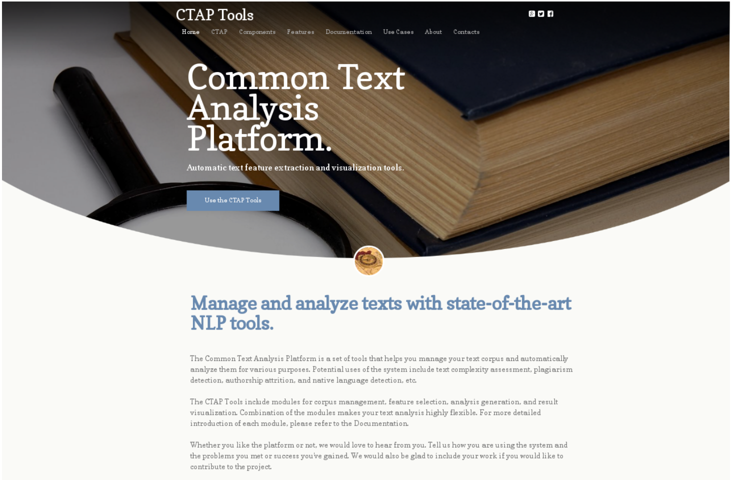


The Result Visualizer is a simple and intuitive module that plots analysis results for the user to visualize preliminary findings from the analysis.

Design features of CTAP

- Consistent, easy-to-use, friendly user interface
- Modularized, reusable, and collaborative development of analysis components
- Flexible corpus and feature management

System demo



CTAP Tools
Home CTAP Components Features Documentation Use Cases About Contacts
Use the CTAP Tools

Common Text Analysis Platform.

Automatic text feature extraction and visualization tools.

Manage and analyze texts with state-of-the-art NLP tools.

The Common Text Analysis Platform is a set of tools that helps you manage your text corpus and automatically analyze them for various purposes. Potential uses of the system include text complexity assessment, plagiarism detection, authorship attrition, and native language detection, etc.

The CTAP Tools include modules for corpus management, feature selection, analysis generation, and result visualization. Combination of the modules makes your text analysis highly flexible. For more detailed introduction of each module, please refer to the Documentation.

Whether you like the platform or not, we would love to hear from you. Tell us how you are using the system and the problems you met or success you've gained. We would also be glad to include your work if you would like to contribute to the project.

http://ctapweb.com

Outlook

- Populating the system with more features
- Replicating studies that involved text analysis to validate the system and identify other function needs
- Model construction functionality (machine learning)
- Acurracy measures
- API supporting analysis of multiple languages (en, de, es, fr...), non-plain text file formats, etc.

More details available in the paper:

Chen, X. B., & Meurers, D. (2016). CTAP: A Web-based tool supporting automatic complexity analysis. In D. Brunato, F. Dell'Orletta, G. Venturi, T. François, & P. Blache (Eds.), *Proceedings of the Computational Linguistics for Linguistic Complexity Workshop at the 26th International Conference on Computational Linguistics (COLING 2016)*, Osaka, Japan, 11 December (pp. 113-119). The International Committee on Computational Linguisitcs.

References

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., and Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2):540–545.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016a). Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, pages 1–19.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., Singh, A., Shankar, Y., Pulickal, L., Rajkumar, A., and Parmar, N. J. (2016). Tacit: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, pages 1–10.

Housen, A. (2015). L2 complexity—a difficult(y) matter. Oral presentation given at the Measuring Linguistic Complexity: A Multidisciplinary Perspective workshop, Université catholique de Louvain, Louvain-la-Neuve.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, Cambridge, M.A.

Vajjala, S. (2015). *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. PhD thesis, University of Tübingen.