EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Xiaobin Chen and Detmar Meurers

LEAD Graduate School & Research Network

LEAD Graduate School & Research Network

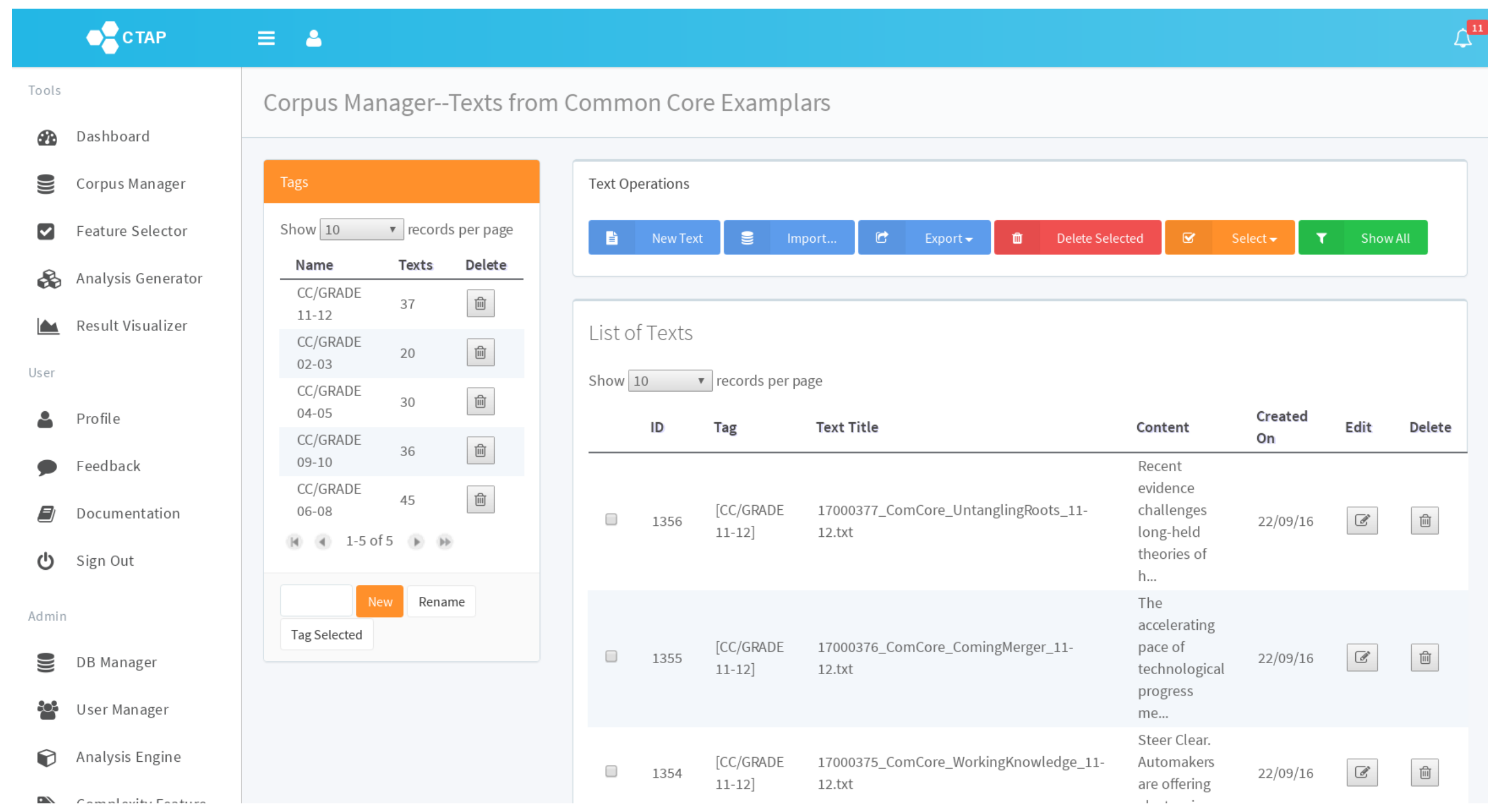
# CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis

## Introduction

- Use of linguistic complexity analysis:
  - assessing text readability
  - modeling processing difficulty of human language
  - assessing second language writing
  - comparing language typologies and their historical development
- Process of complexity analysis:
  - identify the observable variedness and elaborateness (Rescher, 1998; Ellis, 2003) from text
  - interpret results considering reading or writing tasks and the characteristics of the reader or writer
- Focus of the project: a computational system for quantitatively identifying variedness and elaborateness, or *absolute complexity* (Kusters, 2008).

## Existing Tools for Complexity Analysis

- There is a general lack of adequate computational tools for automatic complexity measurement (Bulté and Housen, 2012).
- Limited support for collaborative research, flexible feature management, or cross-platform operationability.
- Limited transparency of the working mechanisms and little extendability from commercial systems.
- Significant amount of feature overlap resulting in waste of precious research resources.
- Hard to use for non-expert computer users.



Corpus Manager--Texts from Common Core Examplars

Show 10 records per page

| Name | Texts | Delete |
|---|---|---|
| CC/GRADE 11-12 | 37 | |
| CC/GRADE 02-03 | 20 | |
| CC/GRADE 04-05 | 30 | |
| CC/GRADE 09-10 | 36 | |
| CC/GRADE 06-08 | 45 | |

1-5 of 5

List of Texts

Show 10 records per page

| | ID | Tag | Text Title | Content | Created On | Edit | Delete |
|---|---|---|---|---|---|---|---|
| □ | 1356 | [CC/GRADE 11-12] | 17000377_ComCore_UntanglingRoots_11-12.txt | Recent evidence challenges long-held theories of h... | 22/09/16 | | |
| □ | 1355 | [CC/GRADE 11-12] | 17000376_ComCore_ComingMerger_11-12.txt | The accelerating pace of technological progress me... | 22/09/16 | | |
| □ | 1354 | [CC/GRADE 11-12] | 17000375_ComCore_WorkingKnowledge_11-12.txt | Steer Clear. Automakers are offering | 22/09/16 | | |

*Figure 1. Corpus Manager Screen Shot*

## CTAP System Architecture

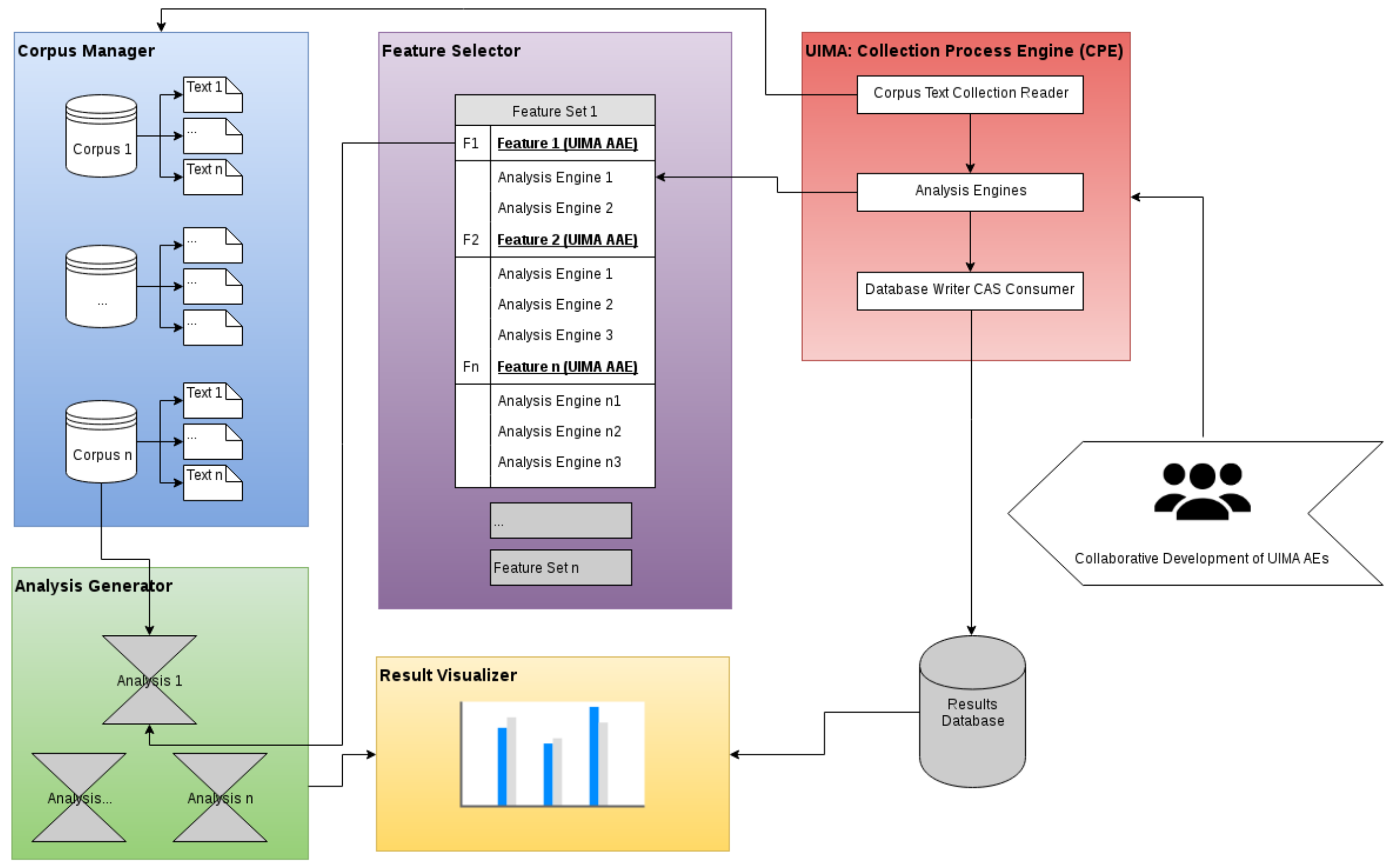


*Figure 2. CTAP System Architecture*

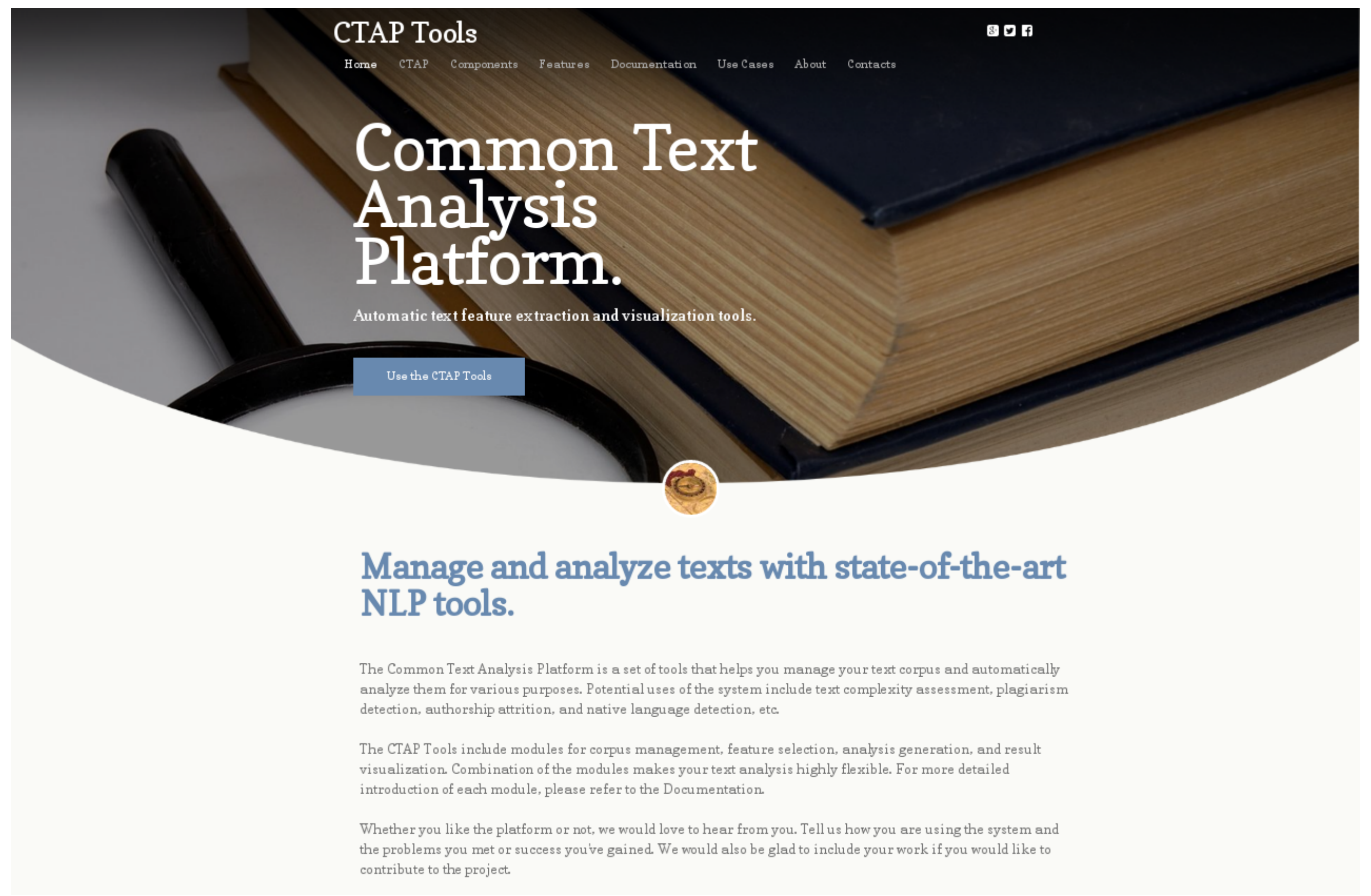
- **Corpus Manager:** helps users manage language materials to be analyzed. It uses
  - folders to group corpora
  - corpora to hold texts
  - tags to label and group texts based on e.g. document genre, target reader levels, etc.
- **Feature Selector:** for selecting complexity features to be extracted from texts. It supports:
  - creating feature set to hold selected features
  - adding/removing features from feature sets
- **Analysis Generator:** extracts a set of features from the designated corpus. It can be used to:
  - create new analyses
  - run analyses and monitor their progress
  - export analysis results
- **Result Visualizer:** a simple and intuitive module that plots analysis results for the user to visualize preliminary findings from the analysis.

## Design Features

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

## Accessing CTAP

`http://ctapweb.com`



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.

Collaboration through `https://github.com/ctapweb`

## Outlook

- Populate the system with additional feature extractors
- Validate and exemplify the system by replicating previous complexity studies (e.g. Vajjala and Meurers, 2012; Hancke et al., 2012)
- Support for multi-language analysis
- Additional functionalities such as team collaboration and statistical modeling

## References

Bulté, B. and Housen, A. (2012). Defining and operationalising l2 complexity. In Housen, A., Kuiken, F., and Vedder, I., editors, *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, pages 21–46. John Benjamins, Amsterdam.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press, Oxford, UK.

Hancke, J., Meurers, D., and Vajjala, S. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Kusters, W. (2008). Complexity in linguistic theory, language learning and language change. *Language complexity: Typology, contact, change*, pages 3–22.

Rescher, N. (1998). *Complexity: A philosophical overview*. Transaction Publishers, London.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173.