CTAP for Chinese: A Linguistic Complexity Feature Automatic Calculation Platform

Yue Cui[1] Junhui Zhu[1] Liner Yang[1] Xuezhi Fang[1] Xiaobin Chen[2] Yujie Wang[3] Erhong Yang[1]
[1]Beijing Language and Culture University, China
[2]Eberhard Karls Universität Tübingen, Germany
[3]Beijing Jiaotong University, China

Introduction

- The construct of linguistic complexity has been widely used in the research of language learning.
- Several existing Chinese text analysis tools provided few features and are lack of corpus management functions.
- They didn't support downloading analysis reports locally as text files.

CTAP

- The Common Text Analysis Platform (CTAP) (Chen and Meurers, 2016) is a web-based language complexity index automatic extraction tool.
- Advantage: friendly user interface, reusable analysis components and open source.
- CTAP supports Chinese, Italian, German, and English.

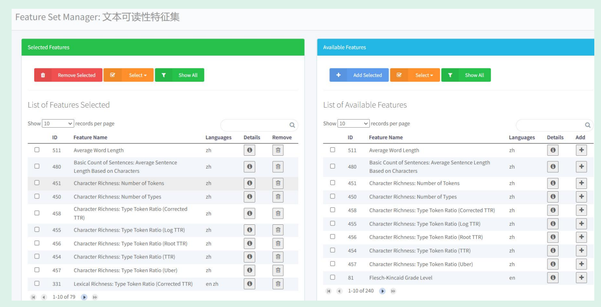
Chinese complexity indexes in CTAP

Level	Subcategory
Character level	Character complexity, Character richness, Character sophistication
Lexical level	Lexical richness, Lexical variation, Lexical density, Lexical sophistication, Word length
Sentence level	Sentence length, Sentence constituent complexity, Syntactic Structure complexity
Paragraph level	Basic count of paragraphs, Cohesive complexity

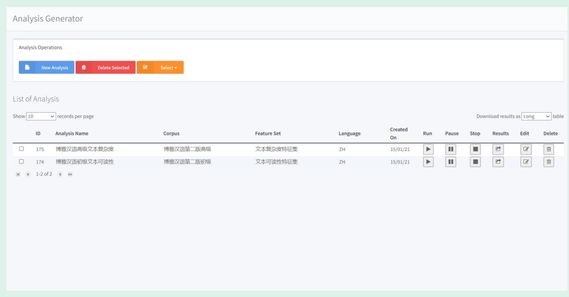
Function of CTAP

Corpus Manger

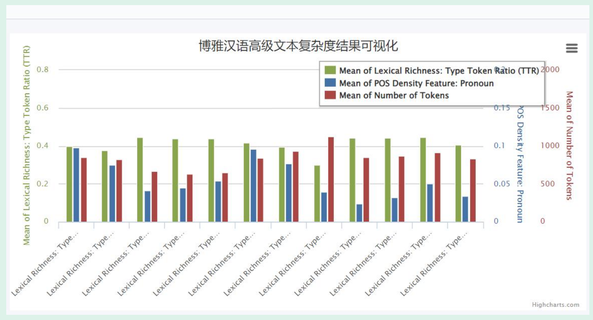
Feature Selector



Analysis Generator



Result Visualizer



Comparing Chinese text analysis tools

		CTAP	CRIE	Chinese Coh-Metrix	Chi-editor
No. of indexes		196	36	50	6
Func-tion	Corpus manager	√	×	×	×
	Index selector	√	√	√	×
	Result visualizer	√	√	×	×
Open source		√	×	×	×
Extendibility		√	×	×	×
Transparency of Results		×	×	×	√

Conclusion

- We constructed a Chinese complexity feature set including 196 features.
- We integrated the Chinese component into CTAP to realize automatic extraction of Chinese complexity features.

E-mail: cyue0829@163.com