

# **CTAP for Chinese: A Linguistic Complexity Feature Automatic Calculation Platform**



#### Introduction

#### **Function of CTAP**

- 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.

Feature Selector

ICALI

Corpus Manager--Texts from 捕班汉语第一版初级

• 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

rTexts from	博雅汉	又语第:	二版初级								Feature	Set N	lanager: 文本可读性特征集							
	Text Op	perations								-	Selecte	d Feature	5				Availa	ble Featu	res	
cords per page	в	New Tex	t 🛢 In	nport 🗈	Export 🗸	Delete Selected	Selec Selec	- 1	Show A	u		Ren	ove Selected 🛛 🐨 Select 🕶 🔽 Show Al	ц., .				<b>ł</b> Ad	Id Selected 🛛 Select 🗸 Show Ali	
IS Delete	Listo	of Texts									List o	f Featu	res Selected			0	List	of Avai	lable Features	
	LISCO	I TEXES	_									ID	Feature Name	Languages	Details	Remove		ID	Feature Name	La
	Show	10	records per pa	ige								511	Average Word Length	zh	0			511	Average Word Length	zh
		ID	Тад	Text Title	Content		Create	d Edi	lit Dele	te		480	Basic Count of Sentences: Average Sentence Length Based on Characters	zh	0			480	Basic Count of Sentences: Average Sentence Length Based on Characters	zł
Rename					大卫・玛丽	雨 那是谁的书? 是你的	)书			_		451	Character Richness: Number of Tokens	zh	0			451	Character Richness: Number of Tokens	zh
		1508	[初级起步	3.那是你的	吗? 玛丽:	不是,那是我同屋的书	方。 方。 15/01/1	21	2			450	Character Richness: Number of Types	zh	0			450	Character Richness: Number of Types	zh
			扁山	书吗?.txt	大卫: 是汉	又语课本						458	Character Richness: Type Token Ratio (Corrected	zh	0	đ		458	Character Richness: Type Token Ratio (Corrected TTR)	zh
			初级起步	2 你是哪国	刘老师: 同	同学们好! 学生: 老师好	仔!					455	(Log TTP)	zh	6			455	Character Richness: Type Token Ratio (Log TTR)	zh
		1507	篇 I ]	人.txt	刘老师:我来介绍一下儿。我姓刘,	J, NJ 15/01/	21	۲. E			455	Character Richness: Type Token Ratio (Log Trik)	zh	0			456	Character Richness: Type Token Ratio (Root TTR)	zh	
					XJ明, 是你						0	450	Character Richness: Type Token Ratio (ROOT TTR)	211	0			454	Character Richness: Type Token Ratio (TTR)	zh
		1506	[初级起步	1 亿加子 tyt	大卫: 你对	f! 李车: 你好! 大卫: 本安· 不具 我不具ま	: 你 判而 15/01/	1 6	7			454	Character Richness: Type Token Ratio (TTR)	211	0			457	Character Richness: Type Token Ratio (Uber)	zł
	0	1000	篇 I ]	1.10.73.101	我是学		57/P7 15/01/					331	Lexical Richness: Type Token Ratio (Oper)	en zh	6			81	Flesch-Kincaid Grade Level	er

阔 🖪 1-10 of 79 🕟 😥

#### Analysis Generator

📢 🖪 1-3 of 3 🕩 🗰

#### Result Visualizer



## **Comparing Chinese text analysis tools**



ΠЪ	<b>511</b>	21	1.
$\sim$			

### **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						

		CTAP	CRIE	Coh- Metrix	Chi- editor
No. c	of indexes	196	36	50	6
	Corpus manager		X	X	X
Func -tion	Index selector				X
	Result visualizer			X	X
Ope	n source		×	X	×
Exte	endibility		X	X	X
Tran of	sparency Results	X	X	X	

Sentence level	Sentence length, Sentence constituent complexity, Syntactic Structure complexity	• We incl	
Paragraph level	Basic count of paragraphs, Cohesive complexity	<ul> <li>We</li> <li>to r</li> <li>con</li> </ul>	



- 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.