CTAP for Chinese: A linguistic Complexity Feature Automatic Calculation Platform

Yue Cui[123], Junhui Zhu[123], Liner Yang[123*], Xuezhi Fang[123], Xiaobin Chen[4], Yujie Wang[5], Erhong Yang[13]

[1]National Language Resources Monitoring and Research Center Print Media Language Branch, Beijing Language and Culture University, Beijing 100083, China

[2]School of Information Science, Beijing Language and Culture University, Beijing 100083, China

[3]Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China

[4]Hector-Institut für Empirische Bildungsforschung, Tübingen Universität, 72072 Tübingen, Germany

[5]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

cyue829@163.com, {nysyzxzjh, lineryang, jasonfang3900, chxiaobin}@gmail.com, yujie.wang@bjtu.edu.cn, yerhong@blcu.edu.cn

Abstract

The construct of linguistic complexity has been widely used in the research of language learning. Several text analysis tools have been made to automatically analyze linguistic complexity. However, the indexes supported by several existing Chinese text analysis tools are limited and varied due to different research purposes. CTAP is an open-source toolkit for linguistic complexity measurement extraction, which serves all research purposes. Although it was originally developed for English, the Unstructured Information Management (UIMA) framework it used allows the integration of other languages. In this study, we integrated the Chinese component into CTAP, describing the index sets it incorporated and comparing it with three linguistic complexity tools for Chinese. The index set includes 4 levels of 196 linguistic complexity indexes: character level, word level, sentence level, and discourse level. So far, CTAP has implemented automatic calculation of complexity characteristics for four languages, aiming to help linguists without NLP background do their research on language complexity.

Keywords: linguistic complexity analyzer, Chinese, text analysis

1. Introduction

Linguistic complexity is a multifaceted construct used in a wide range of contexts (Bult and Housen, 2014), manifested in the variety and complexity of production units or grammatical structures (Wolfe-Quintero et al., 1998). It has been used in various studies, such as second language proficiency and development assessment (Crossley and McNamara, 2014; Bult and Housen, 2014; Kyle, 2016), readability assessment (Vajjala and Meurers, 2012; Feng et al., 2010; Chen and Meurers, 2018), and first language academic writing (Crossley et al., 2011; Weiss and Meurers, 2019). A large number of complexity indexes have been proposed in the above research tasks. However, a common and crucial question these studies have to answer is to what extent the complexity indexes that have existed are valid, which would directly bear upon the validity of the research results. The best way to solve this problem is to compare the proposed indexes with a large number of texts of varying difficulty (Lu, 2010), which requires reliable computational tools that can automatically calculate linguistic complexity index values.

Most of the linguistic complexity work has focused on English (Okinina et al., 2020), so many computational tools can automatically analyze English texts, such as Coh-Metrix, D-Level Analyzer, L2 Syntactic Complexity Analyzer, CTAP, and so on. However, the scope of complexity research has been broadened towards the Chinese in recent years. There are several automated text complexity analysis tools for Chinese, such as the Chinese Readability Index Explorer (CRIE) (Sung et al., 2016), Chinese Coh-Metrix, and Chi-Editor (Jin et al., 2018). Among the complexity indexes provided by these tools, some are different due to their different research purposes, which will be pointed out in Section 2. However, they use different technologies that ultimately lead to different values even for the same indexes. In addition, these tools did not provide corpus management functions or support users in downloading analysis reports locally as text files.

For the above reasons, we integrated Chinese components into the Common Text Analysis Platform (CTAP) (Chen and Meurers, 2016) to support the analysis of Chinese texts. CTAP is a language complexity feature extraction platform for English. The unstructured information management framework it uses makes it extendable, facilitating the addition of new language components. More importantly, it is open-source. So far, the platform has supported the analysis of the complexity indexes for four languages: English, German, Italian and Chinese. The effectiveness and plausibility of textual indexes vary between languages due to their linguistic peculiarities (Sung et al., 2016), so we only transferred 40 linguistic complexity indexes provided for English, German and Italian by integrating a Chinese text processing tool and added 150 indexes for

* Corresponding author: Liner Yang.

Chinese. The Chinese component of CTAP provides a more comprehensive set of indexes than the other existing computational tools for Chinese texts, with support of extractions for 196 linguistic complexity measures, among which average dependency distance, maximum dependency distance, average syntax tree height, and the maximum syntax tree height we implemented are not only applicable to Chinese but also English to measure text complexity.

In this article, we describe the Chinese component of CTAP. The rest of the paper is organized as follows: Section 2 illustrates Chinese complexity research and Chinese automated textual-analysis tools existing. Section 3 introduces the functions of CTAP and NLP tools used in the Chinese component and lists the Chinese complexity indexes it supports. Section 4 describes different characteristics by comparing CTAP and other linguistic complexity measurement tools for Chinese. Finally, the last section concludes the paper and discusses the scope for future work.

2. Related Work

In recent years, with the continuous advancement of information technology, natural language processing, computational linguistics, and second language acquisition, other related disciplines have continued to developed continuously. Researchers are paying more and more attention to the field of text complexity. The study of English complexity started earlier and has produced many research achievements. However, in recent years, the scope of complexity research has been broadened towards Chinese. Taking the field of L2 learning as an example, the research includes investigating the syntactic complexity of the learner's L2 production (Wu, 2016a), the longitudinal development of L2 linguistic complexity (Wu, 2016b), the relationship between complexity and L2 writing quality (Wu, 2018), the effects of learning tasks on the complexity of the learner's L2 production (Wu and Hu, 2021)and comparison of the complexity features in different genres in writing by second language learners (Wu, 2019). In addition, complexity has been used to assess text readability. A Chinese complexity index system with 165 specific complexity indexes was constructed, including four levels of Chinese characters, vocabulary, syntax, and text to assess Chinese textbooks' readability (Wu, 2020). 85 complexity indexes were used to make a refined readability assessment of second language teaching materials (Zhu, 2020). The analysis and calculation of so many indexes must rely on the help of computers, but this is a challenge for researchers without technical background. Therefore, it is necessary to develop an automatic extraction tool for text complexity indexes.

To the best of our knowledge, there are several computational systems for automatic linguistic complexity analysis of Chinese texts. For example, CRIE was designed to analyze text complexity and readability. It is composed of three systems: CRIE, CRIE-CFL (Chinese as a foreign language), and CRIE-DK (Domain Knowledge). Among the three systems, CRIE was used to analyze Chinese texts for native speakers, CRIE-CFL was created to analyze foreign-language reading materials for CFL, and CRIE-DK was invented to assess the knowledge content levels of texts, such as the readability and conceptual difficulty of a web page or e-book (Sung et al., 2016). Chinese Coh-Metrix was created to analyze the cohesion and coherence of Chinese texts. Therefore it focuses on referential cohesion measures (e.g., Local Nouns Overlap), connective measures (e.g., Coordinating Connective), and latent semantic analysis measures for Chinese. It was developed based on Coh-Metrix (Graesser et al., 2011), an online tool for assessing English texts coherence, often used in second language acquisition. Chi-Editor was invented to analyze foreign-language reading text readability, aiming to offer difficulty level of reading material to the international Chinese teacher.

None of the above tools are specifically used to calculate and extract complexity features, so the complexity features they implement are not comprehensive. By adding the Chinese components to CTAP, we aim to provide a complexity index extraction tool with a broader index set than the existing three Chinese complexity analysis tools. CTAP calculated many features about Chinese character components, syntax trees, dependency information, collocation information, and grammar level to make the text complexity assessment more accurate. In addition, CTAP provides corpus management and visualization functions and support to download the calculated index values to the local in the text form , which is convenient for researchers to use in subsequent research. Finally, CTAP allows to aggregate new indexes with its flexible and extendible architecture.

3. CTAP and Its Extension to Chinese

3.1. CTAP Architecture

The Common Text Analysis Platform (CTAP) (Chen and Meurers, 2016) is a web-based language complexity index automatic extraction tool. This tool is not limited to specific research tasks, and the unstructured information management framework it uses is very convenient for adding new indexes or other language modules. In this section, we will make a brief introduction to CTAP as well as the NLP tool we use for Chinese text preprocessing.

3.1.1. An Overview of CTAP User Modules

The CTAP system consists of four basic modules: Corpus Manager, Feature Selector, Analysis Generator, and Result Visualizer. Corpus Manager enables users to upload many texts and build multiple corpora. Feature Selector is used to group the selection of the complexity indexes into index sets. Then, users can utilize Analysis Generator to generate analyses. Result Visualizer is a module for visualizing these analysis results. These modules are designed to be user-friendly and powerful, which enable users to focus more on research findings instead of paying too much effort to text preprocessing and annotation procedures. By the way, the system is compatible with Chinese components as it is language-independent.

3.1.2. NLP Tools for Text Preprocessing

To annotate Chinese texts with high quality in an automatic way, we decide to choose the Stanford CoreNLP toolkit for annotations. CoreNLP is a widely-use, high-performance, and multi-lingual NLP tool created by the Stanford NLP Group (Manning et al., 2014). It contains a pipeline to produce a set of annotations with high precision given a raw text, which we use for sentence splitting, tokenization, part-of-speech tagging, constituency parsing, and dependency parsing. The annotated text is used to extract indexes later on.

3.2. Complexity Indexes for Chinese in CTAP

At present, the Chinese component of CTAP includes 196 linguistic complexity indexes, 40 of which are available for English, German and Italian. These common indexes contain lexical richness (e.g., type-token ratio), lexical variation (e.g., verb variation), lexical density (e.g., noun density), number of syntactic constituents (e.g., verb phrase), and syntactic complexity (e.g., mean length of noun phrases). Because there are large morphological and syntactic differences between the characteristics of Chinese and alphabetic languages (Sung et al., 2016), 150 complexity indexes included the Chinese component of CTAP are unique. The implemented measures are distributed among the following four levels: character level, lexical level, sentence level, and paragraph level.

3.2.1. Character Level

Character level can be divided into three subcategories: character complexity (10 indexes), character richness (9 indexes), and character sophistication (14 indexes).

Character complexity Chinese characters are composed of components, while components are composed of strokes. The number of components and strokes are valid indexes to measure character complexity. Wang and Peng (1999) proposed that there is a significant stroke number effect in the processing of Chinese characters (Wang and Peng, 1999). The more strokes, the slower the processing speed. For this reason, we developed two types of indexes: one is stroke-count-based (e.g., the average number of character strokes) and the other is component-count-based (e.g., the average number of Chinese character components). Besides, we divided Chinese characters into three categories according to the number of strokes: Low-stroke-count characters, Intermediate-stroke-count characters, and High-stroke-count characters. Low-stroke-count character contains 1 to 8 strokes, Intermediate-stroke-count character contains 9 to 16 strokes and High-stroke-count character contains more than 16 strokes (Wu, 2020). Information about strokes is from the Chinese Proficiency Test Application Form Word (2006), which is developed by the Ministry of Education and State Language Commission and contains 5,500 characters. Information about components is from Network of the Chinese characters research team [1], which includes 4,033 characters.

Character richness Character richness refers to the degree of variation of Chinese characters used in a text. There are many ways to calculate the richness of Chinese characters. The most widely used indicator is the Type-Token Ratio (TTR), which is the ratio of the type of Chinese characters in the text to the total number of characters in the text (Templin, 1957). The higher the value of TTR, the more abundant Chinese characters are used. However, the validity of the TTR indicator will be affected by the length of the text. Therefore, we added Log TTR, Root TTR, Uber TTR, and Corrected TTR to measure the character richness of a text. Besides, characters that appear only once in a text are seen as one of the signs that distinguish a text from other texts and can reflect on character richness (Islam et al., 2012). Thus, we also use the number of characters that appear only once and the proportion of characters that appear only once to measure character richness.

Character sophistication The sophistication of Chinese characters is affected by the frequency of Chinese characters in daily life. The higher the frequency of Chinese characters, the easier it is to recognize them. The character sophistication indexes are calculated by referring to the Chinese character frequency tables, which are the Chinese Gigaword character frequency table, L2 Chinese Textbooks character frequency table, and Contemporary Chinese Corpus character frequency table. The Chinese Gigaword Corpus [2] is is a comprehensive archive of newswire text data that has been acquired from Chinese news sources. The L2 Chinese Textbooks Corpus contains many textbooks, which can reflect the usage of each character in educational settings. The Contemporary Chinese Corpus is a large-scale balanced corpus, which contains about 20 million characters. There are two types of indexes in the character sophistication category: Logarithmic Character Frequency (Type) and Logarithmic Character Frequency (Token).

In the field of Chinese second language teaching and acquisition, the difficulty level of Chinese characters stipulated in the Chinese Proficiency Grading Standards for International Chinese Language Education (referred to as the Grade Standard later) are important factors to measure the difficulty of Chinese characters. The higher the level of Chinese characters specified in the Grade Standard, the more difficult it is.

[1] https://learnm.org

[2] https://catalog.ldc.upenn.edu/LDC2011T13

3.2.2. Lexical Level

The lexical level mainly involves five subcategories: lexical richness (9 indexes), lexical variation (9 indexes), lexical density (28 indexes), lexical sophistication (26 indexes), and basic count of words (9 indexes).

Lexical richness The lexical richness indexes refer to the degree of repetition of words used in a text which regardless of the part of speech. The fewer words that are repeated in a text, the richer the words used. Same as the character richness, the lexical richness indexes include: TTR, Log TTR, Root TTR, Corrected TTR, Uber TTR, number of words appearing only once, and proportion of words appearing only once.

Lexical variation The lexical variation indexes reflect the degree of variety of five content word types: nouns, verbs, adjectives, adverbs, and all content words. Content words play an important role in conveying the information of the article. Research shows that people would spend more time processing sentences with more content words (Carpenter and Just, 1983). These indexes calculate the ratio of the number of the five content word types to the number of all content word tokens. The content words contain nouns, verbs, adjectives, adverbs, numerals, measure words, pronouns, distinguish words, interjections, and onomatopoeic words (Huang and Liao, 2011). Verbs often serve as the core component of sentences, so we gave it the special attention that not only the ratio with the number of all content word tokens is considered, but also the ratio with the number of verb tokens.

Lexical density The lexical density indexes refer to the density of every part of speech. In Chinese, all words are divided into content words and function words. Content words are used to convey information and express meaning, while function words are used to organize the structure of the text and express logical relationships. The more content words in the text, the greater the vocabulary density and the greater the amount of information conveyed (Johansson, 2008). Existing studies have also found that function words have a great contribution to the prediction of text difficulty (Sung et al., 2015; Wang, 2005; Wang, 2017; Zuo and Zhu, 2014). Therefore, we add the density indexes of various content words and function words to CTAP. The function words contain prepositions, conjunctions, auxiliary words, and sentence-final particles (Huang and Liao, 2011). The lexical density indexes calculate the ratio of the number of different parts of speech to the number of word tokens.

Lexical sophistication Lexical sophistication can be measured by word frequency. Many studies have indicated that word frequency is related to word response time, with participants responding faster to words that appear more often (Forster and Chambers, 1973; Whaley, 1978). The lexical sophistication indexes are calculated separately for all words, lexical words, and functional words, and each of them is based on the Chinese Gigaword, L2 Chinese Textbooks Corpus, and Contemporary Chinese Corpus. In addition, it introduces the vocabulary level information in the Grade Standard, which calculates the difficulty of vocabulary by mean value and variance of grade, as well as the proportion of simple words (1-3) and difficult words (7-9).

Word length The last subcategory is word length. There is an inverse relationship between lexical length and lexical frequency. The longer the word, the lower the frequency used in the text (Deng and Feng, 2013). Based on the above analysis, we can draw a conclusion that the longer the word, the more difficult it is for people to understand it. Thus, we calculated the number of single-character words, two-character words, three-character words, and four and more characters words.

3.2.3. Sentence Level

The sentence level mainly involves the analysis of three aspects of the sentence in texts: sentence length (8 indexes), sentence constituent complexity (18 indexes), and syntactic structure complexity (28 indexes).

Sentence length The sentence length indexes include the mean sentence length in characters and words, its standard deviation, and the longest sentence length. Generally, the longer the sentence, the greater the amount of information and the more difficult it is to understand the sentence.

Sentence constituent complexity Sentence constituent complexity indexes include the number of every syntactic constituent, its mean length, and diversity. The sentence constituents mainly include noun phrases, verb phrases, prepositional phrases, coordinate phrases, adjectival modifiers, and sentences. The denser the syntactic structure, the higher the cognitive burden on readers. Existing studies have found that the number of noun phrases, verb phrases, and adjective phrases in a sentence is related to the grade of Chinese textbooks for elementary and middle schools (Sun, 2015; Jiang, 2018). Thus, we developed the number of every sentence constituent, its mean length, and diversity to measure sentence constituent complexity.

Syntactic structure complexity The syntactic structure complexity is mainly analyzed from the following four aspects: the parse tree depth, the dependency distance, the grammar, and the collocation. The height of the parse tree can effectively reflect the complexity of the syntax. It has been proven that the higher the parse tree, the more complex the sentence (Sun, 2015; Jiang, 2018; Wu and Hu, 2021). We counted the distribution of syntax trees height in the Chinese Treebank 8.0 and took 80% of the distribution (that is 14) as the boundary value. Based on this, we measure the syntactic complexity by using mean parse tree depth, its standard deviation, the height of the highest parse tree, the number of sentences whose parse tree height is greater than 14, and their proportion. In addition, dependence distance can reflect the cognitive difficulty of sentences (Liu, 2007). Dependence distance refers to the linear distance between words with syntactic relations. The research results of cognitive linguistics show that in syntactic processing, the linear distance between two syntactically related words affects the storage and integration of working memory. As the linear distance between two words becomes longer, the cognitive cost also increases (Gibson and others, 2000). Therefore, indexes about dependence distance include: the average and the maximum number of words before the head verb, mean dependence distance, and maximum dependence distance. Grammar level is also an important factor affecting text comprehension for Chinese second language learners. Indexes about grammar include the average of grammatical levels, the proportion of grammar at each level, and the RTTR of grammar at each level. Hu (2021) proposed collocation-based features of syntactic complexity.

3.2.4. Paragraph Level

The paragraph level mainly involves two subcategories: basic count of paragraphs (5 indexes) and cohesive complexity (23 indexes).

Basic count of paragraphs The basic count of paragraphs indexes include the number of paragraphs, mean paragraph length, and the longest paragraph length.

Cohesive complexity The cohesion was divided into lexical cohesion, reference and logical cohesion (Cai, 2020), which we use to analyze cohesive complexity. Lexical cohesion mainly refers to the pattern of lexical repetition and cohesion in discourse, which includes local cohesion and global cohesion. Local cohesion refers to the lexical repetition between close-distance clauses, while global cohesion refers to the lexical repetition between far-apart sentences (Hoey, 1991).

The lexical cohesion is calculated separately for all words, content words, nouns, and verbs, including cohesion between adjacent sentences and cohesion among all sentences in a text. The reference contains the proportion of the first person pronouns, the second person pronouns, the third person pronouns, interrogative pronouns, and demonstrative pronouns. The logical cohesion includes the proportion of coordinating connectives, alternative connectives, progressive connectives, condition connectives, hypothetical connectives, causal connectives, purposive connectives, and concessive connectives.

4. Comparing Linguistic Complexity Analysis Tools for Chinese

In the following we compare four linguistic complexity analysis tools for Chinese: CRIE, Chinese Coh-Metrix, Chi-Editor and CTAP, describing the main differences among them. These differences are analyzed from the following dimensions: First, we compare index sets in the different tools. Secondly, we present different functions of the four tools. Then we describe the interpretation of results calculated by the tools and their source code availability and we discuss the tools' extendibility and the transparency of the intermediate analysis steps at last. Table 1 presents an overview of the comparison.

4.1. Linguistic Complexity Indexes

Because the research aims of the four tools are different, the index sets they provide are also different. CRIE focuses on the complexity of native language texts, second language texts, and texts in specific domains, so it provides 36 indexes, including 29 general indexes and 7 indexes that are only applicable to the complexity of second language texts. Chinese Coh-Metrix focuses on the cohesion of the text, so among the 50 indexes it provides, 31 indexes are used to measure the cohesion of the article. Chi-Editor aims to grade the reading materials provided by Chinese international teachers, so it will give the level of second language reading materials, but only 6 indicators are analyzed. However, CTAP is not limited to specific research goals, so it provides more generic and comprehensive index sets, including 196 indexes. Since Chi-Editor provides fewer features, it is not discussed in this part of the comparison.

Only 13 complexity indexes are present in all three tools, including character complexity (e.g., low-stroke-count characters), lexical richness (e.g., type-token ratio), POS density index (e.g., pronoun), basic count of sentences (e.g., mean sentence length in tokens) and cohesive complexity index (e.g., personal pronouns per token). The vast majority of measures are different. we will analyze the differences in these indexes from four aspects: character level, lexical level, sentence level, and paragraph level. Table 2 gives a detailed comparison of the indexes provided by these tools.

4.1.1. Character Level

For character complexity, CRIE and Chinese Coh-Metrix provide indexes related to the number of strokes while CTAP uses the number of components and number of characters that appear only once in addition to the number of strokes. For character richness, CTAP offers more fine-grained character richness indexes than CRIE, such as Root TTR, and Uber TTR. For character sophistication, CTAP is the only tool that uses reference corpora to measure character sophistication.

4.1.2. Lexical Level

For lexical richness, CRIE only uses 4 indexes to measure lexical richness included the number of tokens, type-token ratio, the total number of content words, and the total number of negation words. Chinese Coh-Metrix use 7 indexes, including type-token ratio, content words type-token ratio, the measure of textual lexical diversity (MTLD), the number of tokens, the number of word types, the number of content words types, and the number of content words tokens. Apart from the two indexes, CTAP offers various variants of TTR. Chinese Coh-Metrix provides the measure of textual lexical diversity (MTLD) and content word type-token

		CTAP	CRIE	Chinese coh-metrix	Chi-editor
No. of indexes		196	36	50	6
Function	Corpus manager	yes	no	no	no
	index selector	yes	yes	yes	no
	Result visualizer	yes	yes	no	no
Source code availability		open source	proprietary	proprietary	proprietary
Extendibility		extendible	not extendible	not extendible	not extendible
Transparency of Results		no	no	no	yes

Table 1: Comparison of CTAP, CRIE, Chinese coh-metrix, and Chi-editor

ratio. Among these features, CTAP does not provide indexes about content words, negative word count, and MTLD, but it provides various variants of TTR to reduce the impact of text length.

For lexical variation, CTAP is the only tool offering lexical variation indexes.

For lexical density, CTAP analyzed the density of 9 types of content words and 4 types of function words, while CRIE only analyzed the density of three parts of speech, and Chinese Coh-Metrix analyzed the density of seven parts of speech.

For lexical sophistication, the three tools introduce different reference corpora. Since CTAP is not limited to specific research goals, it employs two native corpora: the Chinese Gigaword and Contemporary Chinese Corpus. In addition, it also introduces L2 Chinese Textbooks Corpus to serve Chinese international teachers. CRIE uses the list of 8000 Chinese words published by the Steering Committee for the Test of Proficiency-Huayu (SC-TOP) (Chang, 2012) to measure the lexical sophistication of learners' reading materials. Chinese Coh-Metrix uses Children corpus to measure children reading materials' lexical sophistication.

For word length, the three tools calculate the number of two-character and three-character words. Based on this, the Chinese Coh-Metrix also calculates the number of words with four or more characters. CTAP calculates the number of words with four or more characters and the number of single-character words.

4.1.3. Sentence Level

For sentence length, CTAP not only calculates the average sentence length and the longest sentence length but also calculates the standard deviation of the sentence length to analyze the distribution of sentence length. CRIE only calculated the average sentence length and the longest sentence length.

For sentence constituent complexity, CTAP pays more attention to the number of various phrases (e.g., noun phrases), their mean length and density, while CRIE only supports the calculation of the number of sentences and idioms and the density of noun phrases, idioms, and simple sentences.

For syntactic structure complexity, in particular, CTAP is the only tool that introduces syntax trees, dependency, grammar, as well as collocation information to measure the complexity of syntax.

4.1.4. Paragraph Level

For the basic count of paragraphs, CTAP provides 5 indexes about the number and length of paragraphs, while CRIE only provides the number of paragraphs and the mean number of sentences per paragraph.

For cohesive complexity, only 3 indexes of cohesive complexity have been implemented in CRIE. CTAP offers many indexes about reference, connection, and word overlap to measure cohesive complexity. Chinese Coh-Metrix apart from these indexes also uses sentence syntax similarity, minimal edit distance, and LSA.

4.2. Function

4.2.1. Corpus Manager

During the texts upload process, Coh-Metrix and Chi-Editor only support uploading one text at a time through the web page. CRIE provides two ways for users to upload texts: one is to upload a text directly on the web page, and the other is to upload multiple texts at a time in packages, but it should be noted that the size of the packages must not exceed 15M. CTAP provides corpus management functions, supporting uploading, storage, and management of multi-text corpus so that researchers can analyze the same batch of texts from different perspectives.

4.2.2. Feature Selector

In the feature selection process, since the main purpose of Chi-Editor is to provide international Chinese teachers with the level of difficulty in reading text, it does not support feature selection but gives six default index values as a reference after text analysis. Both Chinese Coh-Metrix and CRIE allow users to select indexes. However, the Chinese Coh-Metrix only supports selecting indexes in one dimension (e.g. coherence) at a time. CRIE allows users to select multi-dimensional indexes but does not support saving the selected index set. CTAP supports the construction of index sets, users can customize multiple index sets according to different research purposes and content. In addition, each index has a detailed explanation and source, which is convenient for users to choose.

4.2.3. Result Visualizer

The four tools have different ways of presenting analyze results. The Chi-Editor and Chinese Coh-Metrix directly display indexes values on the web page, while the analysis report provided by CRIE is presented in a visual form, including the complexity index values and the corresponding level. CTAP displays the complexity index values of the text or corpus input by the user and does not compare these values with the preset reference values. This supports researchers to use their corpus and indexes flexibly to achieve different research purposes. In addition, CTAP not only supports downloading the analysis results directly to the local but also supports basic drawing operations to visualize the results, so as to better serve the user's next analysis.

4.3. Interpretation of Results

CRIE provides a reference standard for users. It tells the users whether the index values contained in the article are higher or lower than the index values of different levels of articles in the reference corpus. In other words, it can tell the users difficulty level of the input article based on every single index.

Chi-Editor uses "The Graded Chinese Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages" (Liu, 2010) and "International Curriculum for Chinese Language Education" (CLEC, 2008)as the grading reference standards, and generates text difficulty values through algorithms and provides directly difficulty level of the text.

Chinese Coh-Metrix and CTAP only provide the value of each index but do not provide specific readability estimates based on reference corpora or external reference data. It purposefully reports only numerical index values, aiming to give users the greatest degree of freedom, allowing users to compare with any corpus that is suitable for research purposes.

4.4. Source Code Availability

Among the existing linguistic complexity analysis tools for Chinese, only CTAP is open-source. CRIE, Chinese Coh-Metrix, and Chi-Editor only provide a web-based graphical interface for users to use online. CTAP not only provides a free and open online version but also provides open-source code to encourage relevant researchers to add new indexes or new languages to jointly promote the development of language complexity research. The Chinese component of CTAP is available open-source at `https://github.com/blcuicall/multilingual-ctap`. The toolkit is freely available at `https://ctap.litmind.ink`.

4.5. Extendibility

Among the three Chinese Textual-analysis tools, Chinese Coh-Metrix, CRIE, and Chi-Editor only support the analysis of Chinese text. CTAP is developed for multilingual collaborative research and can be extended to other languages. At the same time, the various parts of the CTAP framework are independent of each other, facilitating index updates and integration of new languages.

4.6. Transparency of Results

Word segmentation and part-of-speech tagging are the basis of all index calculations. It is important for researchers to see the specific results of word segmentation and part-of-speech tagging. Among the four text complexity analysis tools, only Chi-Editor gives the specific results of word segmentation.

5. Conclusion and Future Work

With the deepening of language complexity research, more and more indexes have been proposed. How to extract, calculate and verify the effectiveness of these indexes on a large scale has become an important issue. In this paper, we have integrated the Chinese component which supports broad index extraction into CTAP. So far, CTAP supports the extraction of 196 indexes to analyze Chinese texts. Among the 196 indexes, indexes related to syntax trees (e.g., mean parse tree depth index)and dependency structures (e.g., mean dependency distance)can be used for English. In addition, we used component information with Chinese character characteristics to measure the complexity of Chinese characters (e.g., the average number of character Strokes). At the same time, we have made a comprehensive comparison between CTAP and the three existing language complexity analysis tools for Chinese. CTAP not only supports comprehensive language complexity index extraction but also allows adding new complexity indexes and language components.

In the future, we will incorporate semantic indexes and add more indexes with Chinese characteristics, such as grammatical structure, etc. We will present the segmentation and pos tagging results of Stanford CoreNLP on the platform and allow users to check them manually. Then CTAP will automatically calculate the feature values with the proofread results to ensure the accuracy of the feature values and the transparency of the results. We hope that more languages can be integrated into CTAP through cooperation to promote the development of language complexity research.

6. Acknowledgements

This work was supported by the funds of Beijing Advanced Innovation Center for Language Resources (No. TYZ19005), Research Project of the National Language Commission (No. ZDI135-131, No. ZDI135-105) and Fundamental Research Funds for the Central Universities in BLCU (No. 21PT04).

7. Bibliographical References

Bult, B. and Housen, A. (2014). Conceptualizing and measuring short-term changes in l2 writing complexity. *Journal of second language writing*, 26:42–65.

Cai, J. (2020). *Research on L2 Chinese Text Readability*. Ph.D. thesis, Beijing Language and Culture University.

Carpenter, P. and Just, M. (1983). What your eyes do while your mind is readinzg. In *Eye movements in reading*, pages 275–307. Elsevier.

Chang, L. (2012). The study of the vocabulary size at the cefr levels for cfl/csl learners. *Journal of Chinese Language Teaching*, 9(2):77–96.

Chen, X. and Meurers, D. (2016). Ctap: A web-based tool supporting automatic complexity analysis. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pages 113–119.

Chen, X. and Meurers, D. (2018). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

CLEC. (2008). *International Curriculum for Chinese Language Education*. Foreign Language Teaching and Research Press.

Crossley, S. and McNamara, D. (2014). Does writing development equal writing quality? a computational investigation of syntactic complexity in l2 learners. *Journal of Second Language Writing*, 26:66–79.

Crossley, S., Weston, J., McLain Sullivan, S., and McNamara, D. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28(3):282–311.

Deng, Y. and Feng, Z. (2013). A quantitative linguistic study on the relationship between word length and word frequency. *Journal of Foreign Languages*, (03):31–41.

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A comparison of features for automatic readability assessment.

Forster, K. and Chambers, S. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, 12(6):627–635.

Gibson, E. et al. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Graesser, A., McNamara, D., and Kulikowich, J. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press.

Hu, R. (2021). On the relationship between collocation-based syntactic complexity and chinese second language writing. *Applied Linguistics*, (1):13.

Huang, B. and Liao, X. (2011). *Modern Chinese*. Higher Education Press.

Islam, Z., Mehler, A., and Rahman, R. (2012). Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553.

Jiang, Z. (2018). Research on text representation technologies for readability assessment. *Nanjing University*.

Jin, T., Lu, X., Guo, K., and Li, B. (2018). Eng-editor: An online english text evaluation and adaptation system. *Guangzhou: LanguageData (languagedata.net/tester)*.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53:61–79.

Kyle, K. (2016). Measuring syntactic development in l2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics*, 15:1–12.

Liu, Y. (2010). *The Graded Chinese Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages*. Beijing Language and Culture University Press.

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

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. pages 55–60.

Okinina, N., Frey, J., and Weiss, Z. (2020). Ctap for italian: Integrating components for the analysis of italian into a multilingual linguistic complexity analysis tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7123–7131.

Sun, G. (2015). Research on readability prediction methods based on linear regression for chinese documents. *Nanjing University*.

Sung, Y., Lin, W., Dyson, S., Chang, K., and Chen, Y. (2015). Leveling l2 texts through readability: Combining multilevel linguistic features with the cefr. *The Modern Language Journal*, 99(2):371–391.

Sung, Y., Chang, T., Lin, W., Hsieh, K., and Chang, K. (2016). Crie: An automated analyzer for chinese texts. *Behavior research methods*, 48(4):1238–1251.

Templin, M. (1957). Certain language skills in children; their development and interrelationships.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Wang, C. and Peng, D. (1999). The roles of surface frequenches,cumulative morpheme frequencies,and semantic transparencies in the processing of compound words. *Acta Psychologica Sinica*, (03):266–273.

Wang, L. (2005). *Research on Chinese Readability Formula of Texts for Elementary and Intermediate*

Korean and Japanese Students. Ph.D. thesis, Beijing Language and Culture University.

Wang, L. (2017). Research on chinese readability formula of texts for elementary and intermediate south korean and japanese learners. *Language Teaching and Linguistic Studies*, No.187(05):15–25.

Weiss, Z. and Meurers, D. (2019). Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 380–393.

Whaley, C. (1978). Word—nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, 17(2):143–154.

Wolfe-Quintero, K., Inagaki, S., and Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Number 17. University of Hawaii Press.

Wu, J. and Hu, R. (2021). Effect of task complexity on the argumentative writing of csl learners. *Chinese Language Learning*, (2):9.

Wu, J. (2016a). The grammatical complexity in english native speakers' chinese writing. *Language Teaching and Linguistic Studies*, No.180(04):27–35.

Wu, J. (2016b). Research on lexical richness development in csl writing by english native speakers. *Chinese Teaching in the World*, 030(001):129–142.

Wu, J. (2018). The research of indices of the grammatical complexity in south korean native speakers' chinese writing quality. *Linguistic Sciences*, 17(05):66–75.

Wu, J. (2019). Comparative study of linguistic features in different genres in chinese writing by south korean student. *Language Teaching and Linguistic Studies*, No.199(005):1–12.

Wu, S. (2020). Automatic evaluation of chinese text readability based on multi-level linguistic features. Master's thesis, Beijing Language and Culture University.

Zhu, S. (2020). *Research on Automatic Assessment of Chinese Text Readability for Second Language Teaching*. Ph.D. thesis, Beijing Normal University.

Zuo, H. and Zhu, Y. (2014). Research on chinese readability formula of texts for intermediate level european and american students. *Chinese Teaching in the World*, 028(002):263–276.

8. Appendix

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Character Richness			
Type Token Ratio (TTR)	+	-	-
Number of Tokens	+	+	-
Number of Types	+	-	-
Type Token Ratio (Log TTR)	+	-	-
Type Token Ratio (Root TTR)	+	-	-
Type Token Ratio (Uber)	+	-	-
Type Token Ratio (Corrected TTR)	+	-	-
Number of Characters that Appear Only Once	+	-	-
The Proportion of Characters that Appear Only Once	+	-	-
Character Complexity			
Average Number of Character Strokes	+	+	+
Average Number of Character Type Strokes	+	-	-
Low-Stroke-Count Characters	+	+	+
Low-Stroke-Proportion Characters	+	-	-
Intermediate-Stroke-Count Characters	+	+	+
Intermediate-Stroke-Proportion Characters	+	-	-
High-Stroke-Count Characters	+	+	+
High-Stroke-Proportion Characters	+	-	-
Average Number of Chinese Character Components (Based on Character Type)	+	-	-
Average Number of Chinese Character Components (Based on Character Token)	+	-	-
Character Sophistication			
Gigaword Logarithmic Character Frequency (Type)	+	-	-
Gigaword Logarithmic Character Frequency (Token)	+	-	-

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Corpus of L2 Chinese Textbooks Logarithmic Character Frequency (Type)	+	-	-
Corpus of L2 Chinese Textbook Logarithmic Character Frequencys (Token)	+	-	-
Contemporary Chinese Corpus Logarithmic Character Frequency (Type)	+	-	-
Contemporary Chinese Corpus Logarithmic Character Frequency (Token)	+	-	-
Average of Character Levels (Token)	+	-	-
SD Character Levels (Token)	+	-	-
Average of Character Levels (Type)	+	-	-
SD Character Levels (Type)	+	-	-
The Proportion of the Difficult Characters (Token)	+	-	-
The Proportion of the Difficult Characters (Type)	+	-	-
The Proportion of the Simple Characters (Token)	+	-	-
The Proportion of the Simple Characters (Type)	+	-	-
Lexical Richness			
Type Token Ratio (Uber)	+	-	-
Type Token Ratio (Corrected TTR)	+	-	-
Type Token Ratio (TTR)	+	+	+
Type Token Ratio (Root TTR)	+	-	-
Type Token Ratio (Log TTR)	+	-	-
Content Words Type Token Ratio	-	-	+
Measure of Textual Lexical Diversity (MTLD)	-	-	+
Number of Tokens	+	+	+
Number of Word Types	+	-	+

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Number of Content Words Types	-	-	+
Number of Content Words Tokens	-	-	+
Total Number of Content Words	-	+	-
Number of Words that Appear Only Once	+	-	-
The Proportion of Words that Appear Only Once	+	-	-
Total Number of Negation Words	-	+	-
Lexical Variation Feature			
Noun	+	-	-
Squared Verb Variation 1	+	-	-
Corrected Verb Variation 1	+	-	-
Verb	+	-	-
Adverb	+	-	-
Adjective	+	-	-
Verb Variation 1	+	-	-
Modifier	+	-	-
Lexical	+	-	-
POS Density Feature			
Noun	+	-	+
Interjection	+	-	-
Adjective	+	-	+
Punctuation	+	-	-
Ordinal Number ofber	+	-	-
Functional Words	+	-	-
Lexical Words	+	+	+
Cardinal Number ofber	+	-	-
Adverb	+	-	+
Preposition	+	-	-
Pronoun	+	+	+
Verb	+	-	+
Conjunction	+	+	+
Number oferal	+	-	-
Measure word	+	-	-
Localizer	+	-	-
Auxiliary Words	+	-	-
Sentence-final particle	+	-	-
Bei-construction	+	-	-
Onomatopoeia	+	-	-
Ba-construction	+	-	-
Average Number of POS tags per sentence	+	-	-
Dummy Verb	+	-	-
Modal Verb	+	-	-
Personal Pronoun	+	-	-
Interrogative Pronoun	+	-	-
Demonstrative Pronoun	+	-	-
Proper Noun	+	-	-

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Lexical Sophistication Feature			
Gigaword Logarithmic Word Frequency (AW Type)	+	-	-
Gigaword Logarithmic Word Frequency (AW Token)	+	-	-
Gigaword Logarithmic Word Frequency (LW Type)	+	-	-
Gigaword Logarithmic Word Frequency (LW Token)	+	-	-
Gigaword Logarithmic Word Frequency (FW Type)	+	-	-
Gigaword Logarithmic Word Frequency (FW Token)	+	-	-
Corpus of L2 Chinese Textbooks Logarithmic Word Frequency (AW Type)	+	-	-
Corpus of L2 Chinese Textbooks Logarithmic Word Frequency (AW Token)	+	-	-
Corpus of L2 Chinese Textbooks Logarithmic Word Frequency (LW Type)	+	-	-
Corpus of L2 Chinese Textbooks Logarithmic Word Frequency (LW Token)	+	-	-
Corpus of L2 Chinese Textbooks Logarithmic Word Frequency (FW Type)	+	-	-
Corpus of L2 Chinese Textbooks Logarithmic Word Frequency (FW Token)	+	-	-
Contemporary Chinese Corpus Logarithmic Word Frequency (AW Type)	+	-	-

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Contemporary Chinese Corpus Logarithmic Word Frequency (AW Token)	+	-	-
Contemporary Chinese Corpus Logarithmic Word Frequency (LW Type)	+	-	-
Contemporary Chinese Corpus Logarithmic Word Frequency (LW Token)	+	-	-
Contemporary Chinese Corpus Logarithmic Word Frequency (FW Type)	+	-	-
Contemporary Chinese Corpus Logarithmic Word Frequency (FW Token)	+	-	-
Average of Word Levels (Token)	+	-	-
SD Word Levels (Token)	+	-	-
Average of Word Levels (Type)	+	-	-
SD Word Levels (Type)	+	-	-
The Proportion of the Difficult Words (Token)	+	-	-
The Proportion of the Difficult Words (Type)	+	-	-
The Proportion of the Simple Words (Token)	+	-	-
The Proportion of the Simple Words (Type)	+	-	-
Breakthrough Vocabulary	-	+	-
Waystage Vocabulary	-	+	-
Threshold Vocabulary	-	+	-
Vantage Vocabulary	-	+	-
Effective Operational Proficiency Vocabulary	-	+	-
Difficult Words	-	+	-
Average of Vocabulary Levels	-	+	-
Mean Square of Vocabulary Levels	-	+	-
Polysemy	-	-	+
Age of Acquisition for All Words	-	-	+
Concreteness for Content Words	-	-	+

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Content Word Frequency	-	-	+
Word Frequency	-	-	+
Minimum Word Frequency per Sentence	-	-	+
Average Logarithmic Frequency of Domain Content Words	-	+	-
Word Length			
Number of Single Character Words	+	-	-
Proportion of Single Character Words	+	-	-
Number of Two Characters Words	+	+	+
Number of Three Characters Words	+	+	+
Proportion of Two Characters Words	+	-	-
Proportion of Three Characters Words	+	-	-
Number of Words that Contains Four Characters and More	+	-	+
Proportion of Words that Contains Four Characters and More	+	-	-
Average Word Length	+	-	-
Syntactic Constituents Complexity Feature			
Noun Phrase	+	-	-
Prepositional Phrase	+	-	-
Coordinate Phrases	+	-	-
Verb Phrase	+	-	-
Adjectival Modifier	+	-	-
Sentences	+	+	-
Idioms	-	+	-
Mean Length of Noun Phrase (token)	+	-	-
Mean Length of Prepositional Phrase	+	-	-
Mean Length of Verb Phrase (token)	+	-	-
Prepositional Phrases per Sentence	+	-	-
Noun Phrases per Sentence	+	+	-
Coordinations per Sentence	+	-	-
Number of Simple Clause Per Sentence	+	-	-
Mean Number of Noun Phrase Per Simple Clause	+	-	-

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Mean Number of Verb Phrase Per Simple Clause	+	-	-
Mean Number of Coordinate Phrase Per Simple Clause	+	-	-
Modifiers per Noun Phrase	-	+	-
Verb Phrases per Sentence	+	-	-
Mean Number of Prepositional Phrases Per Simple Clause	+	-	-
Simple Sentence Ratio	-	+	-
Idioms per Sentence	-	+	-
Sentence Length Feature			
Average Sentence Length Based on Characters	+	-	-
Mean Sentence Length in Tokens	+	+	+
SD Sentence Length (Based on Word Token)	+	-	-
SD Sentence Length (Based on Character Token)	+	-	-
SD Sentence Length (Based on Character Type)	+	-	-
SD Sentence Length (Baesd on Word Type)	+	-	-
Number of Longest Sentence Length (Based on Character Token)	+	-	-
Number of Longest Sentence Length (Based on Word Token)	+	+	-
Syntactic Complexity Feature			
Sentences with Complex Semantic Categories	-	+	-
Mean Parse Tree Depth Feature	+	-	-
SD of Parse Tree Depth	+	-	-
The Height of the Highest Parse Tree	+	-	-
Number of Sentences Whose Parse Tree Height Is Greater Than 14	+	-	-

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Proportion of Sentences Whose Parse Tree Height Is Greater Than 14	+	-	-
Average Number ofber of Words Before the Head Verb	+	-	-
Maximum Number ofber of Words Before the Head verb	+	-	-
Mean Dependency Distance	+	-	-
Maximum Dependency Distance	+	-	-
Average of Grammatical Levels	+	-	-
SD Grammatical Levels	+	-	-
The Proportion of the First-level Grammar	+	-	-
The Proportion of the Second-level Grammar	+	-	-
The Proportion of the Third-level Grammar	+	-	-
The Proportion of the Fourth-level Grammar	+	-	-
The Proportion of the Fifth-level Grammar	+	-	-
The Proportion of the Sixth-level Grammar	+	-	-
First-level Grammar (RTTR)	+	-	-
Second-level Grammar (RTTR)	+	-	-
Third-level Grammar (RTTR)	+	-	-
Fourth-level Grammar (RTTR)	+	-	-
Fifth-level Grammar (RTTR)	+	-	-
Sixth-level Grammar (RTTR)	+	-	-
Total Grammar RTTR	+	-	-
Total Collocation RTTR	+	-	-
Unique Collocation RTTR	+	-	-
UniqueCollocation Ratio	+	-	-
LowfreqCollocation Ratio	+	-	-
Basic Count of Paragraphs			
The Number of Paragraphs	+	+	-

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Mean Number of Sentences per Paragraph	-	+	-
Mean Paragraph Length (Based on Character Token)	+	-	-
Mean Paragraph Length (Based on Word Token)	+	-	-
Longest Paragraph Length (Based on Character Token)	+	-	-
Longest Paragraph Length (Based on Word Token)	+	-	-
Cohesive Complexity Feature			
Personal Pronouns per Token	+	+	+
1st Person Pronouns per Token	+	-	+
2nd Person Pronouns per Token	+	-	+
3rd Person Pronouns per Token	+	-	+
Interrogative Pronouns per Token	+	-	+
Demonstrative Pronouns per Token	+	-	+
Coordinating Conncetives per Token	+	-	+
Follow Conncetives per Token	+	-	+
Alternative Conncetives per Token	+	-	+
Progressive Conncetives per Token	+	-	+
Condition Conncetives per Token	+	-	+
Hypothetical Conncetives per Token	+	-	+
Causal Conncetives per Token	+	-	+
Purposive Conncetives per Token	+	-	+
Concessive Conncetives per Token	+	-	+
Positive Conjunctions	-	+	-
Negative Conjunctions	-	+	-
Local Lexical Words Overlap	+	-	+
Local Nouns Overlap	+	-	+
Local Verbs Overlap	+	-	+
Local Words Overlap	+	-	-
Global Lexical Words Overlap	+	-	+

Linguistic complexity measure	CTAP	CRIE	Coh-metrix
Global Nouns Overlap	+	-	+
Global Verbs Overlap	+	-	+
Global Words Overlap	+	-	-
Sentence Syntax Similarity (Adjacent Sentences)	-	-	+
Sentence Syntax Similarity (across Paragraphs)	-	-	+
Minimal Edit Distance (Part of Speech) (Adjacent Sentences)	-	-	+
Minimal Edit Distance (Part of Speech) (Across Paragraphs)	-	-	+
Minimal Edit Distance (All words) (Adjacent Sentences)	-	-	+
Minimal Edit Distance (All words) (Across Paragraphs)	-	-	+
Local LSA	-	-	+
Global LSA	-	-	+
LSA Given/New	-	-	+
LSA Verb Overlap	-	-	+