The StickyText Program: Automatic Analysis of Text Cohesion and Readability

Xin Rong

Abstract

StickyText is a computer program that assists writing instruction by automatically analyzing cohesion and readability of text, and provides visualized feedback to users. The program supports the evaluation of text cohesion based on statistical language models, graph-based lexical analysis, and various objective readability metrics. Compared with existing software applications that provide similar functions, the StickyText project aims to build a transparent connection between the conventional linguistic accounts of text cohesion and the computational practices. The other major goal of the project is to make the interface more simplistic and user-friendly than similar programs.

In this paper, I provide a brief review of linguistic accounts of text cohesion and readability assessment, and explore the existing related computational models. Then I introduce the StickyText program, showing its design, function and usage. With StickyText, I make some preliminary observations on the cohesion patterns of student writings. I also discuss the future directions of development.

1 Introduction

The assessment of the quality of student-written essays is a crucial component of writing instruction. Instructors usually assign scores to student writings as a general feedback, and provide detailed advice to students helping them improve specific aspects of the writing. In America, the study in text quality assessment has been motivated by the national concern over the poor writing skills of high school graduates during the 1970s and 1980s (Kellogg, 1994). Instead of testing students' writing ability indirectly, such as using multi-choice questions on vocabulary and grammar, direct assessment of a writing sample has been widely adopted. The massive need for reliable and generalizable assessment methods calls for rigorous research in text quality evaluation. Various statistical models and software have been developed to facilitate teachers' assessment of student writing (Witte and Faigley, 1981; Burstein, 2003; Graesser et al., 2004; Nukoolkit et al., 2011).

Kellogg (1994) defines *text quality* to be "judgments about how well a document communicates or achieves its purpose with its intended audience." In this paper, I use the terms *readability* and *text quality* alternatively, which both have the same meaning. There are three categories of methods for measuring text readability, including *subjective measures*, *objective measures*, and *behavior measures. Subjective measures* require human judges to read the writing samples and assign subjective scores based on either their impression or specific properties of the text. *Objective measures* utilize computational methods to analyze text statistics, structure and features to generate readability scores, and provide revision suggestions. *Behavioral measures* usually involve experiments in which human subjects are asked to read the documents and answer comprehension questions or perform other behavioral tasks. These three types of measures are discussed in closer details later.

An important metric in measuring text quality is *text cohesion*. It is concerned with how easily the reader can recover the semantic meanings of a sentence in a context without making much inference. To achieve text cohesion, adjacent sentences should share text elements that connect with each other, namely the *cohesive ties*, so that the workload of comprehension can be reduced for readers. Cohesive ties may be constructed by simple repetition of words, or synonyms, pronouns, conjunctions, and ellipsis. The sentences and paragraphs in a document should also follow the *given-new* contract (Clark and Clark, 1977) to maintain a reader-friendly information flow. Various computational models of text cohesion have been proposed as enhancement of objective readability measures.

Investigating readability and text cohesion measures is very important for facilitating instructors with the writing instruction process. There exists some good attempts, such as the *Coh-Metrix* (Graesser et al., 2004), which provides a rich set of readability and cohesion evaluations given any input document. However, it is difficult for instructors or students to directly utilize the results due to the lack of interpretability of the statistical results. There are other software applications that focus on locating strengths and weaknesses of an essay based on local cohesion and readability metrics. For example, (Nukoolkit et al., 2011) offers a integrated visualized interface that highlights recommended revision places in a document.

The StickyText program is built to combine the advantages of the above programs, aiming to provide comprehensive cohesion and readability metrics, and also make such results intuitively understandable by users.

The rest of the paper is organized as follows. In Section 2, I first review the concept of text cohesion in conventional linguistic literature, and introduce computational models of text cohesion, many of which I have implemented in the program. Then I discuss general issues of text readability assessment, and several existing computer programs that support readability measurement. In Section 3, I officially introduce the *StickyText* program, describing its design, functions and usage, followed by several analyses and observations made with StickyText on student-written essays in Section 4. I end by giving discussions on future direction of StikyText development and general applications of text cohesion analysis.

2 Modeling Text Cohesion

In this section, I review the accounts of text cohesion in conventional linguistic and psychological literature, and then proceed to computational models of text cohesion and readability measures.

2.1 Linguistic Accounts of Cohesion

Definition of cohesion. Hoey (1991) defines *cohesion* as "a property of text whereby certain grammatical or lexical features of the sentences of the text connect them to other sentences in the text." Hoey also provides a categorization of text cohesion, in which he distinguishes lexical cohesion from other types of cohesive ties. Lexical cohesion, according to Hasan (1984) and Hoey (1991), includes repetition, synonymy, antonymy, hyponymy, and meronymy, each of which corresponds to a specific kind of relationship between tokens or words that function together to contribute to cohesion in a certain context. Other kinds of textual ties include reference, conjunction, ellipsis, and substitution.

Cohesion and coherence. In much of the linguistic literature, *coherence* is used as a distinct concept from *cohesion*. De Beaugrande and Dressler (2011) argue that cohesion considers the connections between words, phrases, and sentences on the surface level of text, while coherence is involved with more hidden relationships, such as ties between semantic concepts underlying the superficial sentences. Similarly, in other literature, *cohesion* is described as the phenomenon that the same entities are referred to in successive sentences within a discourse and *coherence* corresponds to the fact that *who*, *what*, *when*, *where*, and *why* remain consistent within a certain range of context (Bishop, 1997; Gernsbacher, 1990; Harley, 2001).

In addition, Hasan (1984) claims that cohesion is a static property of the text, while the perception of coherence is dynamic, which means a single piece of text may be perceived to have multiple levels of cohesion by different individuals. In this paper, I focus on text cohesion, the property of the surface text, instead of coherence, the property of the underlying meanings. However, it would be very meaningful to study how patterns of text cohesion influences readers' perception of coherence.

It is also noteworthy that cohesion occurs on multiple levels of text components, including words, sentences and paragraphs. In addition, text components should not only connect to each other locally, but also contribute to the global development of topics as a whole (Kintsch and Van Dijk, 1978). In addition, Hoey (1991) argues that the meaning of a cohesive text as a whole is greater than the sum of the meanings of its individual components, which illustrates the goal and meaning of addressing cohesion in natural language.

The *Given-New Contract.* The semantic meanings delivered by sentences in a document form a flow of information. Such a flow should satisfy the so-called *given-new* *contract* (Clark and Clark, 1977). This contract regulates the writer's behavior in organizing sentences and paragraphs so that newer information is presented in a context where older information is readily accessible in the reader's short term memory. This usually requires the writer to appropriately repeat or rephrase the older information. The *given-new contract* reduces the cognitive costs for readers to interpret the meanings delivered by the writer. The term, *given-new*, is sometimes also referred to as *topic-comment*, or *theme-rheme*.

The techniques taken by the writer to achieve text cohesion are known as *cohesive devices*. As Kellogg (1994) suggests, *cohesive devices* may include referential ties (e.g., pronominal anaphora), syntactic ties (e.g., using conjunctions), lexical ties (e.g., paraphrasing), and inferential ties (e.g., linking sentences using world knowledge).

A model that addresses the realization of the *given-new contract* is given by the *centering theory* (Grosz et al., 1995; Kruijff-Korbayová and Hajièová, 1997). The theory claims that each utterance in a document has one or multiple forward-looking centers, and a single backward-looking center. The backward-looking center in a subsequent sentence is connected to one of the forward-looking centers in its preceding sentence. The theory then imposes a set of constraints on how the centers in the sentences can be realized. For example, it claims that no forward-looking centers in a previous sentence can be realized as a pronoun in the following sentence, unless the backward-looking center of that following sentence is also realized as a pronoun (Kruijff-Korbayová and Hajièová, 1997).

Despite the wide acceptance in attention modeling and discourse analysis, the *centering theory* encounters great difficulty in applications on large-scale corpora, because a necessary step to apply the analysis is human annotation, which is very time-consuming. In the next subsection, I discuss computational measures that can greatly reduce the cost of human efforts in discourse analysis and text quality assessment.

2.2 Computational Models of Text Cohesion

Conventional discourse analysis usually involves massive manual coding of discourse materials (e.g., coding the centers for *centering theory* analysis), which is time-consuming and requires a considerable amount of training. Automatic techniques can accurately locate a wide variety of linguistic features, including named entities, keywords, repetitive words, synonyms, collocations and colligations, and thus can be used to discover common patterns in good-quality texts that human judges consider as cohesive and coherent. With these objective metrics, it becomes possible to build models of text cohesion that can be easily applied to large-scale corpora.

Here I look at several computational models of text cohesion, many of which are implemented in the StickyText program.

Basic readability measures. These measures are usually functions of the average number of syllables per word, or average number of words per sentence, which are usually not indicative of text cohesion. I include them here as baselines to be compared with specially designed cohesion measures. A widely used readability formula was developed by Flesch (1948):

$$R.E. = 206.835 - 0.846wl - 1.015sl \tag{1}$$

where $R.E.$ is *reading ease*; wl is the number of syllables per 100 words; and sl is the average number of words per sentence. This score usually falls between 0 (hard to read) and 100 (easy to read), although it may fall outside the range.

There is another set of widely-used readability metrics that estimate the educational grade level necessary to understand a document, including FOG, SMOG, and Flesch-Kinciad metrics , among which the Flesch-Kinciad metric is most often used (Si and Callan, 2001). The Flesch-Kinciad grade-level measurement is defined as

$$G.L. = -15.59 + 11.80wl + 0.39sl \tag{2}$$

where $G.L.$ is *grade level.* It can be seen that $G.L.$ uses the same features as $R.E.$, but falls on a different scale.

More sophisticated readability metrics go beyond surface linguistic features, and utilize statistical models learned from linguistic corpora, so that such metrics can be sensitive to the content of the target document (Si and Callan, 2001).

Word-based models. An straightforward idea to measure the cohesion of text is to count the number of overlapping words between adjacent sentences. This idea is based on the assumption that word repetition occurring among adjacent sentences increases the easiness for readers to build semantic connections while reading the text, and hence that piece of text is perceived to be more cohesive. It is defined as

$$W.C. = \frac{1}{N} \sum_i ow(S_i, S_{i+1}) \tag{3}$$

where $W.C.$ is *word-based cohesion*; ow is the number of overlapping words; and N is the number of sentences in the document. In addition, by measuring the average number of overlapping words between every pair of (adjacent and non-adjacent) sentences in a document, we can get a sense of the global cohesion of text. This metric is given by

$$G.W.C. = \frac{2}{N(N-1)} \sum_{i,j} ow(S_i, S_j) \tag{4}$$

where $G.W.C.$ is *global word-based cohesion.*

Word-based models have a clear disadvantage: they cannot handle cohesive ties between variations of words or topics, such as synonyms and antonyms, and hyponyms. Additionally, in practice, the $W.C.$ and $G.W.C$ scores should also be normalized by sentence lengths, to avoid the influence of sentences that are significant longer than average.

Distribution-based models. Distribution-based models measure the average distances between vector representations of adjacent sentences. A vector representation of a sentence is usually the word frequency distribution of the sentence. The distance is usually defined as the cosine of the angel between vectors representing sentence pairs in the high-dimensional word frequency space. A higher average cosine score indicates better cohesion of the text. The equation for this metric is

$$D.C. = \frac{1}{N} \sum_i \cos < S_i, S_j > \tag{5}$$

where $D.C.$ is *distribution-based cohesion* score.

Distributional metrics are flexible in that representations of documents can be manipulated in a variety of ways. As Foltz et al. (1998) suggest, Latent Semantic Analysis (LSA) can be used to obtain the representation of documents on an abstract semantic space, such that variations of words, including synonyms, antonyms, and hyponyms, can be connected. With LSA representations, a pair of adjacent sentences without any common words may still be recognized as cohesive sentences if their words share similar underlying semantics.

To utilize LSA in cohesion analysis, we first need to train an LSA model on large-scale linguistic corpus, so that we could obtain a semantic vector representation of each word in the vocabulary. Thereafter, we can replace the vector representation of a sentence with the weighted sum of the semantic vectors of the words that appear in the sentence, and thus we can use this new vector representation of sentences to apply Equation 5 to obtain the distribution-based cohesion, which is given by

$$LSA.C. = \frac{1}{N} \sum_i \cos < L(S_i), L(S_j) > \tag{6}$$

where $LSA.C.$ is *LSA-based cohesino* score, and $L(S)$ is the LSA-based vector representation of the sentence.[1]

Graph-based models. The above LSA-based cohesion analysis implies that the analysis of word similarity is the basis of distributional-based models of sentences. Besides LSA, there are various other ways to determine word similarity. A widely used approach to word similarity computation is based on semantic relationship or taxonomy encapsulated in a graph of words, such as the WordNet (Miller, 1995; Resnik, 1995; Teich and Fankhauser, 2004). A typical way to compute the similarity between words on a word graph is to find the shortest path from one word to another, and take the distance as the indicator of the similarity between the two words. It is defined as

$$sim(w_1, w_2) = \frac{1}{\min_{p \in P(w_1,w_2)} p(w_1, w_2)} \tag{7}$$

[1] Due to time constraints, the LSA-based metric has yet to be implemented in the program.

where $sim(w_1, w_2)$ is the taxonomy-based similarity of words w_1 and w_2, and p is the length of path from w_1 to w_2. The taxonomy-based similarity may also be realized as the shared information of two concepts according to the information theory (Lin, 1998):

$$sim(w_1, w_2) = \max_{c_1 \in C(w_1), c_2 \in C(w_2)} \{ \max_{c \in S(c_1, c_2)} [-\log p(c)] \} \tag{8}$$

where $S(c_1, c_2)$ is the set of concepts that subsume both concepts c_1 and c_2, and $p(c)$ is the probability of encountering an instance of concept c. Note that each word w may possess multiple senses, and is thus represented by multiple concepts $C(w)$. For details, see (Lin, 1998).

In addition to WordNet, we can also create a graph of words based on their co-occurrence or higher-order co-occurrence from a large corpus, and thus I can use the resulted graph to evaluate word similarity on new dimensions. For example, I may use *point-wise mutual information* (PMI) to measure the similarity between a pair of words, which is given by

$$PMI(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \tag{9}$$

where $p(w_1, w_2)$ indicates the probability of observing words w_1 and w_2 together in a sentence.[2]

The keywords (or the centers in the *center theory*) of sentences are usually nouns or noun phrases. Therefore, when applying the graph-based models, we can first use a part-of-speech tagger to tag sentences, and then apply the models to only nouns and noun phrases between adjacent pairs of sentences in the document.

Other resources. Morris and Hirst (1991) provide a comprehensive review of lexical cohesion analysis. In recent years, there are various other models that have been proposed to account for computational modeling of text cohesion, especially lexical cohesion, such as entity-matrix-based model (Lapata and Barzilay, 2005), and other advanced graph-based models (Gürkök et al., 2008).

2.3 Text Readability Assessment

Understanding the general approach to text readability assessment helps us better understand the research methods in text cohesion. Here I provide a brief summary of Kellogg's (1994) account of readability assessment. As mentioned in Section 1, there are three types of readability measures: *subjective measures*, *objective measures*, and *behavioral measures*.

Subjective measures require human judges to provide subjective scores to the essays. Usually three types of scoring methods are involved: *primary trait scoring* develops highly specific traits given task instructions that define rhetorical situations; *analytical scoring*

[2] PMI-based model has not been implemented in the program.

scores separate dimensions of text quality, such as organization, idea development, vocabulary, and grammar; *holistic scoring* provides general impression and overall judgment to text quality.

Objective measures focus on automatic metrics of readability. The most basic approach is to check various statistics of the text, such as spelling, punctuation, repetition of words, and sentence length. Advanced automatic systems are capable of analyzing structure and substructure of the document, as well as idea organization, cohesive ties, and various other structural features.

Behavioral measures involve experiments where subjects are required to read the essay and answer comprehension questions or perform other related behavioral tasks. Such a approach enables empirical correlation analysis between the various objective text features and the subjective difficulty of comprehension (the underlying assumption is that the more questions the subjects can answer correctly about the document, the higher the readability of the document).

The correlation between subjective judgments and objective metrics of text readability is very complex. One reason is that such a relationship is usually not linear. For example, it will result in lower readability score if the average length of sentences in a document is either too long or too short (Reed, 1989). Other reasons involve human readers' rich world knowledge and their individual difference in interpreting the propositions delivered by the author.

Such complexity between subjective and objective metrics may be partially understood through behavioral experiments, comparing subjects' comprehension of the document given the manipulated version and the original version of text. The manipulation may include deliberately altering text elements, cohesive instruments, or improving poorly-written locations indicated by software, so that researchers may expect to observe subjects' different reactions to the altered and original versions (Britton et al., 1990; Witte and Faigley, 1981).

2.4 Several Existing Programs

There are several existing computer programs that support text cohesion and coherence analysis. The *Coh-Metrix* (Graesser et al., 2004) is a computer tool that measures text readability at different levels, including vocabulary, syntactic composition, meaning, and cohesion, among which, as the authors admit, cohesion and coherence analysis is the most challenging task. *Coh-Metrix* is able to assess the overall cohesion of texts and provide detail scores of various text properties regarding cohesion. The authors claim that they will also develop the Cohesion Gap Identification Tool, or *Coh-GIT*, which will be a computer tool to identify specific places in text where cohesion gaps exist. By far this tool has yet been published.

The biggest advantage of *Coh-Metrix* is that it supports a considerable number of linguistic features that measure the difficulty of comprehension. It covers cohesion analysis on both local and global scope of text, and on both vocabulary (or lexical) and grammar dimensions. Despite the rich set of features that *Col-Metrix* supports[3], it does little to help writers or instructors to easily interpret the statistical scores, especially locating places in the text that result in the loss of cohesion and coherence. As claimed by the authors, such feature is supposed to be implemented in the upcoming tool, *Coh-GIT*.

In contrast to *Coh-Metrix*, there are other programs that support specific locating problematic sentences or clauses. Nukoolkit et al. (2011) introduce a computer program, named *Text Cohesion Visualizer*, that provides visualized interface for writers and instructors to analyze the cohesiveness between sentences and paragraphs of an essay. It is capable of highlighting sentences that it recognizes as "need revision", or visualize the geometric relationship between paragraphs in the semantic space so that the writer can identify any "outlier" paragraph. The whole framework is based on lexical cohesion analysis enhanced by graph-based models using the WordNet. The authors do not report any empirical experiment to show the reliability of the results.

More generally, various text quality assessment software applications exist, many of which are mature enough to match or even outperform human raters' reliability, one of which systems is the *E-rater* used by the Educational Testing Service (ETS) to automatically score analytic writings in standardized exams (Burstein, 2003). The *LightSIDE* system, an open-source machine-learning-backed writing assessment system, is claimed to match the performance of ETS's proprietary *E-rater*[4]. The Hewlett Foundation sponsors the *Automated Student Assessment Prize* (ASAP) competition, in which data scientists and machine learning specialists are encouraged to develop scoring engines that can score student-written essays closely enough to human expert graders. The competition attracted 156 teams in 2012 with $100,000 awards in total.[5]

Note that the focus of this paper is readability assessment, which is not exactly the same as automatic essay scoring. The latter is aimed to match human rater's judgment to the maximum possible degree, whereas our purpose is to provide most useful information for writers to improve their writings.

3 The StickyText Program

In this section, I introduce StickyText, the program that I write to automatically analyze text cohesion and readability. StickyText currently supports three basic readability metrics, and three cohesion metrics. It also supports highlighting sentences it recognizes as most cohesive or incohesive. The program is primarily written in Python, and also has a web-based visualized interface written in PhP. It refers to several external computing packages, including *NLTK 2.0*, and *PyHyphen 2.0.2*. See 3.2 for detailed explanations.

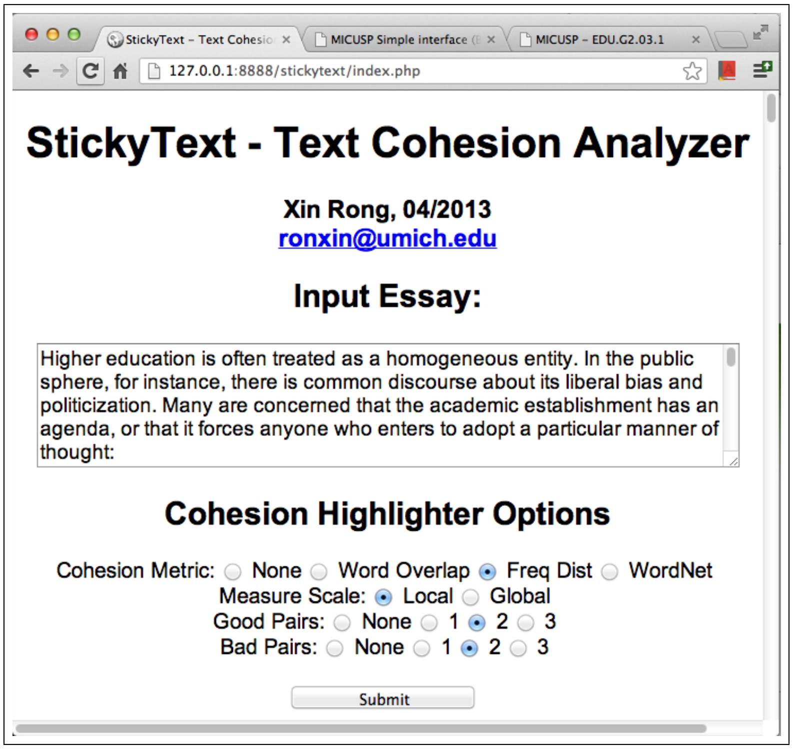
[3] The latest version as of March 2013, *Col-Metrix 3.0*, returns 108 scores for an input document.

[4] According to a talk given by the first author of *LightSIDE*, Elija Mayfield, during his visit to the University of Michigan.

[5] The competition homepage of ASAP: http://www.kaggle.com/c/asap-aes

3.1 Functions

Most of the functions of StickyText can be accessed by the user via a browser. For Python programmers, StickyText can also be imported as a Python module, which enables all of its functions. Figure 1 shows the browser interface of StickyText. In this interface, the user can type in or paste his essay, select the options for cohesion highlighting, and hit the "Submit" button to start cohesion analysis. The result page consists of three sections: (1) readability and cohesion statistics (Figure 2); (2) paragraph-wise cohesion plot (Figure 3); and (3) cohesion annotation (Figure 5).


StickyText – Text Cohesio
MICUSP Simple interface
MICUSP – EDU.G2.03.1
127.0.0.1:8888/stickytext/index.php
Submit

StickyText - Text Cohesion Analyzer

Xin Rong, 04/2013
ronxin@umich.edu

Input Essay:

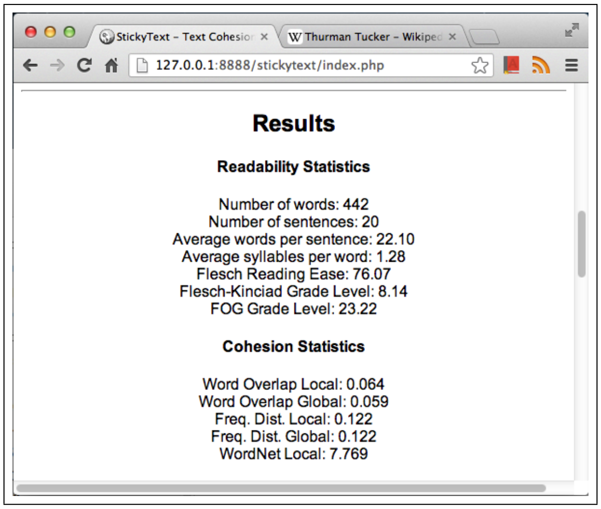
Higher education is often treated as a homogeneous entity. In the public sphere, for instance, there is common discourse about its liberal bias and politicization. Many are concerned that the academic establishment has an agenda, or that it forces anyone who enters to adopt a particular manner of thought:

Cohesion Highlighter Options

Cohesion Metric: ○ None ○ Word Overlap ● Freq Dist ○ WordNet
Measure Scale: ● Local ○ Global
Good Pairs: ○ None ○ 1 ● 2 ○ 3
Bad Pairs: ○ None ○ 1 ● 2 ○ 3

Figure 1: Web interface of StickyText. The user can input essay in plain text form, and select cohesion highlighting options.

Assessing the results of StickyText analysis allows the user to both get a general sense of the reading ease and cohesion of the input essay, and locate specific places in the essay that are most and least cohesive. Based on theses results, a student user may proceed to revise the corresponding sentence, and an instructor user may make respective advice to students. Such an approach, known as revision recommendation, is described in (Kellogg, 1994). It has been used to analyze the effectiveness of objective cohesion metrics by comparing human rater's responses to original essays and essays revised based on recommended


StickyText – Text Cohesio
Thurman Tucker – Wikiped
127.0.0.1:8888/stickytext/index.php

Results

Readability Statistics

Number of words: 442
Number of sentences: 20
Average words per sentence: 22.10
Average syllables per word: 1.28
Flesch Reading Ease: 76.07
Flesch-Kinciad Grade Level: 8.14
FOG Grade Level: 23.22

Cohesion Statistics

Word Overlap Local: 0.064
Word Overlap Global: 0.059
Freq. Dist. Local: 0.122
Freq. Dist. Global: 0.122
WordNet Local: 7.769

Figure 2: Result page of StickyText, showing readability and cohesion statistics on the document level.

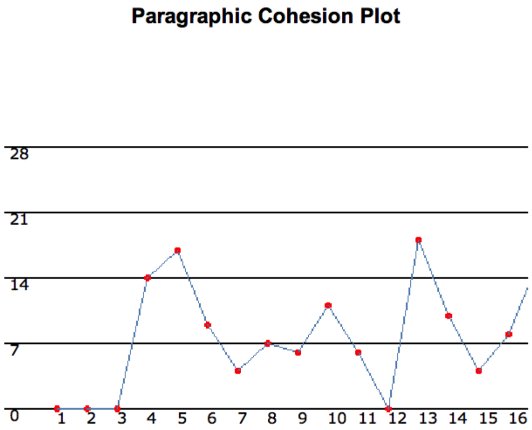


Figure 3: Result of StickyText analysis, showing cohesion scores on the paragraph level. The horizontal axis is paragraph index, and the vertical axis is readability metric selected by the user. The trend of cohesion of an essay can be observed from the graph.

revision locations.

In the future, it will be meaningful to provide cohesion annotation or revision recommendation on a finer scale, such as clauses and words, as well as providing explanation described in natural language about why a specific sentence is highlighted.

3.2 Working Mechanism

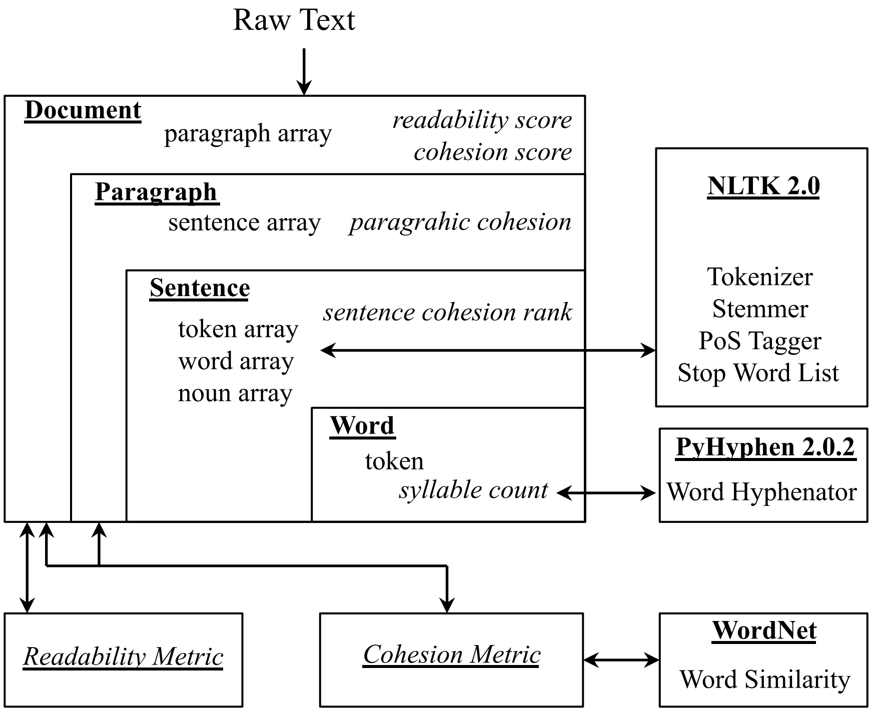


Figure 4: The structure of document representation and modules in StickyText and its external references.

The working mechanism of the StickyText program is shown in Figure 4. When the program takes a new piece of text as input, different parsers parse the text and create a hierarchical document representation internally. As illustrated in Figure 4, a document structure consists of paragraphs; a paragraph structure consists of sentences; and a sentence structure has multiple representations, including a token array (consisting of both words and punctuation, which are useful for certain kinds of analysis such as part-of-speech tagging), a word array, and an entity array (including only noun words); and finally a word structure consists of a token, and its syllables.

The creation of internal document representation utilizes two external packages. *NLTK 2.0*[6] is used for sentence tokenization, word stemming (the Porter stemming algorithm is used), part-of-speech tagging, and stop word removal; and *PyHyphen 2.0.2*[7] is used for determining the number of syllables in a word. Word stemming and stop word removal are important to improve the effectiveness of cohesion metrics based on word overlapping and frequency distribution. Part-of-speech tagging is necessary for retrieving entities (noun words and phrases) from sentences for WordNet-based cohesion metric. The numbers of syllables in words are used in several basic readability metrics.

The three implemented readability metrics are: Flesch reading ease, Flesch-Kinciad grade level, and FOG grade level (see Section 2). And the implemented cohesion metrics include word-based metrics (Eq. 3 and 4), distribution-based metrics (Eq. 5), and WordNet-based metrics (Eq. 7 and 8).

Each of the cohesion metrics are be applied locally (only evaluating cohesion between adjacent sentences) or globally (evaluating cohesion between any pair of sentences in the document). In addition, all cohesion metrics can be evaluated not only on the document level, but also on the paragraph level, so that the user can compare cohesion scores of each individual paragraph in his essay to discover paragraphs that need revision or to observe the general trend of change of cohesion (illustrated in Figure 3).

4 Experiments and Results

To examine the effectiveness of the cohesion metrics as well as the correctness of the implementation, I did some simple experiments with the program using real papers as input. The rest of this section describes the data and the observations from the experiments.

4.1 Data

I selected a small set of essays from the *Michigan Corpus of Upper-Level Student Papers*[8](MICUSP) for testing the StikyText program. To be specific, I selected 10 disciplines from the corpus, including biology, civil engineering, etc. Then, for simplicity, I selected one paper from each discipline. The average number of words per document is 3,159. The longest document is the one in Mechanical Engineering with 13,602 words, and the shortest, Environment, 783 words. For each document, only the main body of text, excluding the title and any references, is used for analysis.

Discipline	Word-based Original	Word-based Shuffled	Dist.-based Original	Dist.-based Shuffled
Biology	0.0318	0.0243	0.0644	0.0505
Civil Engineering	0.0636	0.0284	0.1362	0.0636
Economics	0.0504	0.0180	0.1018	0.0392
Education	0.0691	0.0396	0.1348	0.0769
English	0.0556	0.0234	0.1281	0.0487
History	0.1062	0.0625	0.2176	0.1274
Industrial Eng.	0.0614	0.0220	0.1197	0.0445
Linguistics	0.0690	0.0248	0.1335	0.0487
Mechanical Eng.	0.0791	0.0165	0.1445	0.0306
Environment	0.0374	0.0424	0.0716	0.0777

Table 1: Comparison of cohesion scores before and after random sentence shuffling.

4.2 Results of Cohesion Analysis

Random sentence permutation. As a proof of concept, it is meaningful to see how cohesion scores change when the sentences of a document are altered from their natural order. In theory, the author should try to make text cohesive, and as these papers are all Grade-A papers in college, they should demonstrate good cohesion. Therefore, after random permuting (or "shuffling") the sentences within a document, the cohesion scores should have an apparent decrease from the original value.

Table 1 compares the cohesion scores of the aforementioned ten essays before and after random sentence permutation. Since it takes much longer time for computation, I do not include WordNet-based metrics in this comparison. In addition, all the cohesion scores are evaluated at local scale, which means only the relationship between adjacent sentences are considered in the cohesion measurement (as opposed to at global sale, comparing each pair of sentences within a document).

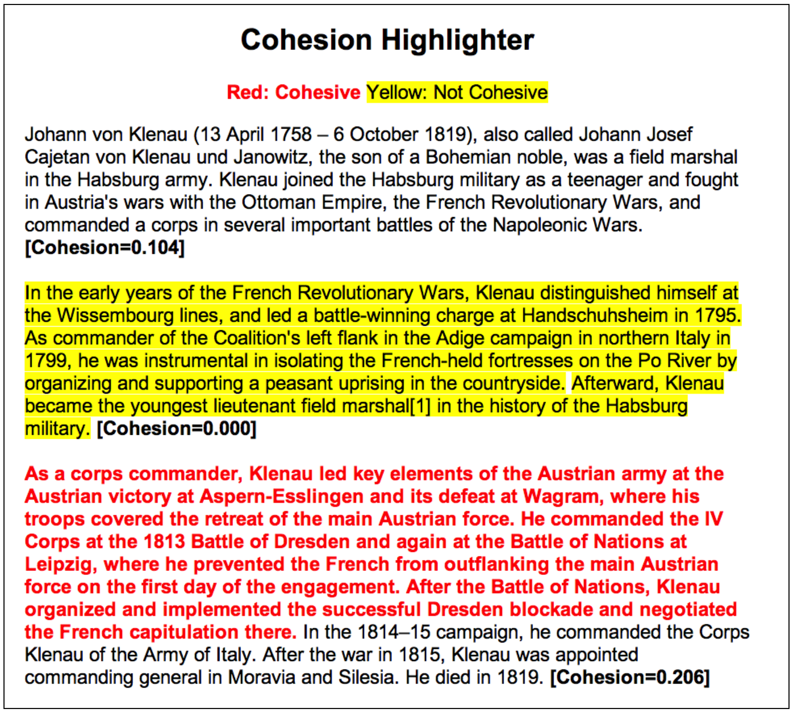
From the results in Table 1, it can be seen that for 9 out of the 10 documents, both metrics exhibit obvious decreases from the original value after random sentence permutation. This result is consistent with our expectation. The paper in Environment is one exception, for which paper both metrics show an increase, instead of decreases. This is possible either because the paper is short and the random effect of our permutation treatment is amplified, or the original paper has poor cohesion and random permutation improves its cohesion "by chance"(which is unlikely).

It is also interesting to compare cohesion scores across disciplines in Table 1, where

[6] The NLTK 2.0 package: http://nltk.org/

[7] The PyHyphen 2.0.2 package: https://pypi.python.org/pypi/PyHyphen/

[8] *MICUSP* homepage: http://micusp.elicorpora.info/



Cohesion Highlighter

Red: Cohesive Yellow: Not Cohesive

Johann von Klenau (13 April 1758 – 6 October 1819), also called Johann Josef Cajetan von Klenau und Janowitz, the son of a Bohemian noble, was a field marshal in the Habsburg army. Klenau joined the Habsburg military as a teenager and fought in Austria's wars with the Ottoman Empire, the French Revolutionary Wars, and commanded a corps in several important battles of the Napoleonic Wars. **[Cohesion=0.104]**

In the early years of the French Revolutionary Wars, Klenau distinguished himself at the Wissembourg lines, and led a battle-winning charge at Handschuhsheim in 1795. As commander of the Coalition's left flank in the Adige campaign in northern Italy in 1799, he was instrumental in isolating the French-held fortresses on the Po River by organizing and supporting a peasant uprising in the countryside. Afterward, Klenau became the youngest lieutenant field marshal[1] in the history of the Habsburg military. **[Cohesion=0.000]**

As a corps commander, Klenau led key elements of the Austrian army at the Austrian victory at Aspern-Esslingen and its defeat at Wagram, where his troops covered the retreat of the main Austrian force. He commanded the IV Corps at the 1813 Battle of Dresden and again at the Battle of Nations at Leipzig, where he prevented the French from outflanking the main Austrian force on the first day of the engagement. After the Battle of Nations, Klenau organized and implemented the successful Dresden blockade and negotiated the French capitulation there. In the 1814–15 campaign, he commanded the Corps Klenau of the Army of Italy. After the war in 1815, Klenau was appointed commanding general in Moravia and Silesia. He died in 1819. **[Cohesion=0.206]**

Figure 5: An illustration of results of StickyText analysis, showing the original document with highlighted sentences. Sentences colored in red are cohesive; sentences colored in yellow are incohesive. The user can change the cohesive metric being applied, and the number of sentences that are annotated. Note that a sentence can be both cohesive and incohesive at the same time (e.g., being cohesive with the preceding sentence, but incohesive with the subsequent one).

one can find that the paper in History exhibits the best cohesion, possibly because of repetition of proper nouns indicating people's names and locations. After random sentence permutation, that paper still has the highest cohesion score, indicating that a number of terms in the paper are repetitively mentioned. However, such an analysis is informal. To obtain concrete conclusions, especially about cohesion and styles in different domains, one needs to look at significantly larger amount of data and takes into account the variance of individual authors, both of which are beyond the scope of the present paper.

Change of cohesion within document. It is very interesting to see how the level of cohesion changes within a document. The visualized interface of StickyText makes it very easy to perform such kind of analysis. I looked at the change of paragraphic cohesion score for each of the ten documents (the results generated by StickText are charts like Figure 3). Not surprisingly, there does not exist a universal pattern for change of cohesion within a document. However, the pattern of changes in subgroups of the documents do exhibit coherence to some degree. For example, many papers have a significant trend of increasing cohesion at the end, possibly because the authors use more cohesive language in discussion and conclusion sections. A small set of papers exhibit a sudden high peak in the middle, which is difficult to interpret.

Highlighted sentences. Another important feature of StickyText is to highlight most cohesive and incohesive pairs of sentences it recognizes in a document. By looking at the highlighted sentences in the results of cohesion analysis, I find that to be highlighted as "most cohesive", a pair of sentences usually need only two common content words (or lemmas). For example, the following pair of sentences are highlighted as the most cohesive pair with distribution-based metric in the paper in Biology, the cohesive ties bolded.

> *The **threat** of bioinvasions must be publicized and **personalized**. Only through the aggregation of countless **personal** efforts can a global **threat** be contained.*

Here is another example showing a pair of sentences recognized as cohesive with distribution-based metric. Note how the author name in the citation contributes to cohesion. This is from a paper in Linguistics:

> *This is the process by which **immigrants** come to sound less like the old **dialect** without sounding like the new **dialect** (**Chambers**, 695). **Chambers** visualizes this process as a continuum with the old and new **dialects** being points and the **immigrant** being somewhere in between.*

StickyText also highlights most incohesive pairs of sentences. However, the simple lexical cohesion metrics it uses to highlight sentences often fire "false negatives", which means they often fail to recognize good cohesive pairs. The following three sentences constitute two adjacent pairs, which are both recognized as most incohesive pairs in the same Linguistic paper:

> *When someone speaks, the listener can tell if he's a local or foreigner and what country he's from. The spread of accents **in Britain** is so wide that within every town, dialects exist. **This** isn't the case **in other English speaking countries** such as the United States and Australia.*

The first and second sentences are indeed not cohesive, by which StickyText is correct. But the second and third sentences are very cohesive. "This" in the third sentence is tied to the wide "spread of accents" described in the second sentence; meanwhile "in other English speaking countries" is tied to "in Britain". However, neither cohesive ties are simple enough to allow StickyText to successfully recognize. More generally, the lexical cohesion metrics cannot recognize cohesion ties based on pronouns, and connections that require world-knowledge to understand (e.g., the fact that Britain is one of the English speaking countries).

5 Discussions

I have shown that StickyText can do basic readability and cohesion analysis on natural language documents, and it also can generate human interpretable results which can be used for revision recommendations or discourse studies. As mentioned above, the function of the program is still very basic and has a number of limitations. First, it can only handle lexical cohesion ties, while in reality other types of cohesive devices are utilized in very flexible ways, such as reference, ellipsis, and substitution. These cohesive devices are usually involved with pronouns, conjunctions, and world knowledge, which require more advanced NLP algorithms to recognize and measure.

Second, StickyText can only evaluate cohesion either between adjacent sentence pairs, or all sentence pairs in a document. In practice, it would also be meaningful to focus on the text-initial sentence (or the topic sentence) and its relationship with all other sentences in the document. Other nested structures of a document should also be taken into account.

Third, StickyText should be able to address cohesion patterns on clause level. For long sentences, the internal ordering of clauses or words can be optimized to improve reading ease. The program should be able to determine the existence of clauses and their connection to any forward-looking center in previous clauses. It is also important to look at semantic roles (e.g., agent, patient, and action) in a sentence and identify any possible connections from them to the contextual sentences.

In sum, it remains a very challenging task to devise automatic techniques that can unveil the nested semantic structure of a natural language essay and capture any cohesion patterns within. The development of such techniques are not only important to automatic essay grading and writing instruction, but also very essential for improving the performance of text summarization, question-answering, and machine-translation, all of which involve generation of natural-language text (Lapata and Barzilay, 2005) where cohesion and readability measures can serve as an evaluation method, or the *likelihood function* in the generative probabilistic models.

References

Bishop, D. V. (1997). *Uncommon understanding: Development and disorders of language comprehension in children.* Psychology Press Hove.

Britton, B. K., Gulgoz, S., and Tidwell, P. (1990). Shaped mental representations of original versus principled revisions of text. In *Psychonomic Society.*

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. *Automated essay scoring: A cross-disciplinary perspective*, pages 113–121.

Clark, H. H. and Clark, E. V. (1977). *Psychology and language.* Cambridge Univ Press.

De Beaugrande, R. and Dressler, W. (2011). Text linguistics. *Handbook of Pragmatics Highlights (HoPH)*, page 286.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Gernsbacher, M. A. (1990). *Language comprehension as structure building.* Psychology Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, &amp; Computers*, 36(2):193–202.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Gürkök, H., Karamuftuoglu, M., and Schaal, M. (2008). A graph based approach to estimating lexical cohesion. In *Proceedings of the second international symposium on Information interaction in context*, pages 35–43. ACM.

Harley, T. A. (2001). *The psychology of language: From data to theory.* Psychology Press.

Hasan, R. (1984). Coherence and cohesive harmony. *Understanding reading comprehension: Cognition, language and the structure of prose*, pages 181–219.

Hoey, M. (1991). *Patterns of lexis in text.* Oxford University Press.

Kellogg, R. T. (1994). *The psychology of writing.* Oxford University Press (New York).

Kintsch, W. and Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5):363–394.

Kruijff-Korbayová, I. and Hajièová, E. (1997). Topics and centers-a comparison of the salience-based approach and the centering theory.

Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In *International Joint Conference On Artificial Intelligence*, volume 19, page 1085. LAWRENCE ERLBAUM ASSOCIATES LTD.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on Machine Learning*, volume 1, pages 296–304. San Francisco.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Nukoolkit, C., Chansripiboon, P., Mongkolnam, P., and Todd, R. (2011). Text cohesion visualizer. In *Computer Science &amp; Education (ICCSE), 2011 6th International Conference on*, pages 205–209. IEEE.

Reed, W. M. (1989). The effectiveness of composing process software: An analysis of writer's helper. *Computers in the Schools*, 6(1-2):67–82.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM.

Teich, E. and Fankhauser, P. (2004). Wordnet for lexical cohesion analysis. In *Proceedings of the Second International WordNet Conference, Brno, Czech Republic.*

Witte, S. P. and Faigley, L. (1981). Coherence, cohesion, and writing quality. *College composition and communication*, 32(2):189–204.