The Cognitive Complexity of Texts and Its Formal Measurement in Cognitive Linguistics

Yingxu Wang1,*, Robert C. Berwick2, and Xiangfeng Luo3

1 International Institute of Cognitive Informatics and Cognitive Computing (ICIC), Department of Electrical and Computer Engineering, Schulich School of Engineering University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada
2 Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
3 Department of Computer Science, Shanghai University, Shanghai 200072, China

In order to rigorously measure the cognitive complexity of texts in natural languages, a formal metric is developed that provides a fundamental measure for the properties of syntax and semantics in textual comprehension, processing, and search. The objective and subjective aspects of text comprehension and their complexity are formally modeled. A formal language model is introduced that characterizes the discourse of natural languages. The mathematical models of cognitive complexity of texts and their properties are rigorously described at the sentence, paragraph, and essay levels from the bottom up. On the basis of the cognitive and mathematical models of cognitive linguistics, the measurement of cognitive complexity of texts is quantitatively established and tested by a set of case studies. A wide range of applications of the measurement of textual complexity are identified in cognitive linguistics and contemporary web technologies such as cognitive search engines, online document retrieval, natural language processing, cognitive linguistics, cognitive machine learning, computing with words, and cognitive computing.


1. INTRODUCTION

Languages are gifted and collective inventions of human beings for thought, expression, communication, as well as knowledge representation and acquisition (Bickerton, 1990; Botha and Knight, 2009; Casti and Karlqvist, 1986; Cavalli-Sforza, 2000; Chomsky, 1982, 2007; Gleason, 1997; Wang, 2009a; Wang and Berwick, 2012; Zadeh, 1999). Linguistics is the science of natural and symbolic languages (Chomsky, 1956; Crystal, 1987; O’Grady and Archibald, 2000; Pattee, 1986; Pullman, 1997; Wardhaugh, 2006). Cognitive linguistics is an emerging discipline that studies the cognitive properties of natural languages and the cognitive models of linguistics in cognitive computing and computational intelligence (Eugene, 1996; Evans and Green, 2006; Langlotz and Andreas, 2006; Taylor, 2002; Wang, 2013; Wang and Berwick, 2012).

It was conventionally perceived that the natural language properties were non-quantifiable and their mathematical modeling was challenging problems in cognitive linguistics and computational linguistics. Studies on formal syntaxes and semantics of natural languages can be traced back to the works of Alfred Taski (1944) and Noam Chomsky (1956). Chomsky formalized a general language framework known as the universal grammar (Chomsky, 1956, 1957, 1959, 1962, 1965, 1982, 2007). Eugene et al. proposed the concept of cognitive linguistics, which attempts to explain the cognitive processes of language and knowledge acquisition, storage, production, and comprehension (Eugene, 1996; Wang and Berwick, 2012). Wang and his colleagues introduce the deductive grammar (Wang, 2009a), deductive semantics (Wang, 2010b) and their denotational mathematical means (Wang, 2008a) in cognitive linguistics and cognitive informatics (Wang, 2002, 2003, 2007a, b, 2009b, c, 2010a; Wang and Wang, 2006; Wang et al., 2006, 2009a,
A concept algebra based technology (Wang, 2008b, 2010c) is developed recently for rigorous semantic modelling and analysis as well as machine learning (Tian et al., 2011; Wang et al., 2011a). Then, new paradigms of denotational mathematics (Wang, 2008a) known as inference algebra and semantic algebra is created for machine reasoning based on language knowledge bases (Wang, 2011a, b, 2013).

Semantic analysis and comprehension are a deductive cognitive process. In cognitive psychology, comprehension is described as the construction of an internal representation based on existing knowledge previously gained in the brain (Matlin, 1998; Polya, 1957). Both Matlin and Polya identified the subjectivity of comprehension that it has a relation with the comprehender’s background knowledge, where whatever one is trying to comprehend, one relies on one’s existing knowledge previously acquired. According to cognitive informatics (Wang, 2010a, 2003, 2007b) and the Object-Attribute-Relation (OAR) model (Wang, 2007c) for internal knowledge representation, the semantics of a
sentence in a natural language is considered being understood only when:
(a) The logical relations of parts of the sentence are clarified; and
(b) All parts of sentence are reduced to the terminal entities, which are either a real-world image or a primitive abstract concept.

In cognitive informatics, cognitive science, AI, and computational intelligence, comprehension is identified as an ability to understand something, which indicates an intelligent power of abstract thought and reasoning of humans or intelligent systems. It is curious to explore the internal process of comprehension in the brain and to explain its basic mechanisms, because comprehension is one of the fundamental processes of brain at the advance cognitive layer according to the Layered Reference Model of the Brain (LRMB) (Wang et al., 2006).

The studies on semantic analyses and text comprehension in cognitive linguistics lead to the analysis of a fundamental problem known as the cognitive complexity of texts and its relation to the syntactic complexity and the semantic complexity in natural languages. On the basis of the formal models and theories of the cognitive complexity of texts, a wide range of problems in cognitive linguistics may be explained. For instances, what is the formal model of the cognitive complexity of texts? What are the roles of and interactions between syntax and semantics in text comprehension? Why may the Internet have significantly reduced the cognitive complexity of texts? What is the threshold that may be used to distinguish a simple and a complex essay? How is the cognitive complexity of texts reduced for certain levels of readers?

This paper presents a quantitative measurement of the cognitive complexity of texts towards the formal explanation of the aforementioned problems. In the remainder of the paper, both objective and subjective aspects of text cognition and their complexity are formally modeled for cognitive linguistics and cognitive computing. Section 2 introduces a cognitive linguistics model of texts. Section 3 analyzes the basic properties of cognitive complexity of texts in natural languages. Section 4 creates a set of mathematical models of cognitive complexity for text comprehension where the metrics of text cognitive complexity is rigorously modeled and measured at the sentence, paragraph, and essay levels from the bottom up. Section 5 quantifies the measurement of text cognitive complexity in comprehensive case studies. A wide range of applications of the measurement of text complexity have been identified in online, web-based, and electronic document comprehension and processing for both humans and cognitive machines in cognitive linguistics and cognitive computing.

2. THE DEDUCTIVE SYNTAX AND SEMANTICS FOR TEXT COMPREHENSION

The structural and semantic properties of natural languages are described in this section based on a generic language model as the discourse of cognitive linguistics and computational linguistics. Then, the formal models of the deductive syntax and semantics are established. The relations and interactions between syntax and semantics of natural languages are explained.

2.1. The Cognitive Linguistic Model of Natural Languages

A formal model of a general language can be described as an abstract language (Wang and Berwick, 2012), which forms the discourse of cognitive linguistics.

**Definition 1.** The abstract language, \( \mathcal{L} \), is a 5-tuple, i.e.,

\[
\mathcal{L} = (\mathcal{A}, \mathcal{M}, \mathcal{S}, \mathcal{N}, \mathcal{D})
\]

where

- \( \mathcal{A} \) is the alphabet of the language as a finite ordered set of letters, \( \alpha_1, \ldots, \alpha_n \), i.e., \( \mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\} \);
- \( \mathcal{N} \), the morphology, is a set of lexical rules between power sets of letters in the given alphabet of \( \mathcal{L} \), i.e., \( \mathcal{N} = (\mathcal{P}\mathcal{A} \times \mathcal{P}\mathcal{A}) \subseteq \mathcal{L} \);
The universal language processing (ULP) model.

Fig. 1. The universal language processing (ULP) model.

2.2. The Deductive Syntax of Cognitive Linguistics

A formal syntax is abstract models of the syntax system in \( V \) where concrete strings of tokens and their grammatical relations are symbolically represented and rigorously analyzed. The syntactic rules that underlie natural languages form the domains of formal linguistics and grammars. One of the most influential linguistic framework known as the theory of universal grammar (UG) was proposed by Noam Chomsky (Chomsky, 1957, 2007). UG is a system of categories, mechanisms, and constraints shared by natural languages, which is perceived as innate based on neurolinguistic and cognitive linguistic studies (Chomsky, 1982; O’Grady and Archibald, 2000; Taylor, 2002).

Definition 2. The mathematical model of syntax, \( \Psi \), in language \( V \) is a Cartesian product of words \( W \), i.e.:

\[
\Psi \equiv W \times W \tag{2}
\]

where \( W \) is a power set of words with legal strings of the alphabet in the language, \( W \subset \mathcal{P}W \subset V \), and \( \Psi \subset \mathcal{P}W \subset V \).

2.3. The Deductive Semantics of Cognitive Linguistics

Semantics is a domain of linguistics that studies the interpretation of words and sentences, and the analyses of their meanings. In linguistics, semantics deals with how the meaning of a sentence in a language is obtained and comprehended. Studies on semantics explore mechanisms in the understanding of language and the nature of meaning where syntactic structures play an important role in the interpretation of sentence and the intension and extension of word meaning (Chomsky, 2007; Taylor, 2002; Wang, 2009a, 2010b, c). A formal semantics is mathematical models for denoting meaning of symbols, notations, concepts, functions, and behaviors, as well as their relations that can be reduced onto a set of predefined entities and/or known concepts.

Definition 3. The mathematical model of semantics, \( \Theta \), in language \( \mathcal{V} \) is a function that maps a set of syntactic relation \( \Psi \) into a set of terminal syntactic relations with terminal words \( \Psi^0 \), i.e.:

\[
\Theta \vdash \Psi \rightarrow \Psi^0 \quad \equiv \quad W \times W \rightarrow \Psi^0 \times \Psi^0 \tag{3}
\]

where \( W \) is a power set of words with legal strings of the alphabet in the language, \( W \subset \mathcal{P}W \subset \mathcal{V} \); and \( \Psi^0 \) is a subset of \( W \) with a power set of terminal words, \( \Psi^0 \subset W \subset \mathcal{P}W \subset \mathcal{V} \).

In cognitive linguistics, the state space of semantics in a natural language can be modeled as follows.

Definition 4. The cognitive model of semantics of natural languages, \( \Theta \), is a 5-tuple, i.e.:

\[
\Theta \equiv (J, V, O, S, T) \tag{4}
\]

where

- \( J \) is the subject of the sentence;
- \( V \) is a behavior or action;
- \( O \) is the object of the sentence;
- \( S \) is the space where the action is occurring;
- \( T \) is the time when the action is occurring.

2.4. Relations Between Syntax and Semantics in Cognitive Linguistics

According to Definitions 1 through 4, the relationship between a language and its syntaxes and semantics can be illustrated as shown in Figure 1 (Wang, 2009a) where the text \( T \) is expressed in one-dimension (1-D), the syntax is in 2-D, and the semantics is in 5-D. Figure 1 explains that linguistic analyses are a deductive process that maps the 1-D text into the 5-D semantics via the 2-D syntactical analyses.
The Cognitive Complexity of Texts and Its Formal Measurement in Cognitive Linguistics

The cognitive informatics model of comprehension is a cognitive process at the advanced cognition layer according to the LRMB model of the brain (Wang et al., 2006) and the object-attribute-relation (OAR) model of internal knowledge representation (Wang, 2007c). The cognitive model of the comprehension process can be conceptually modeled in the following steps:

(a) To search relations from real entities to virtual entities and/or existing objects and attributes
(b) To build a partial or adequate OAR model of the entity.
(c) To wrap up the sub-OAR model by classifying and connecting it to appropriate clusters of the entire OAR in long-term memory.
(d) To memorize the new OAR model and its connections in long-term memory.

The cognitive process of comprehension is supported by the low-level processes such as those of search, memorization, and knowledge representation. A rigorous description of the cognitive process of comprehension may refer to (Wang and Gafurov, 2010).

3. BASIC PROPERTIES OF THE COGNITIVE COMPLEXITY OF TEXTS

Based on the cognitive linguistic models of deductive syntax and semantics in the universal linguistic discourse developed in the preceding section, the basic properties of cognitive complexity of texts can be explained and analyzed. The cognitive complexity of texts and their comprehension is dependent on both its structural and denotational characteristics, i.e., the syntactical and semantic properties of the text. According to cognitive informatics, the former can be measured by the structural complexity of the text; while the latter can be measured by the number of semantic reductions during text comprehension.

**Lemma 1.** The cognitive complexity of texts is proportional to both the syntactic complexity $C_{\Psi}$, dependent on the structural properties of the given text and the semantic complexity $C_{\Theta}$ in term of the number of semantic reductions by searching the meaning of each word in the given text, respectively, i.e.:

$$C \doteq f(C_{\Psi}, C_{\Theta})$$  \hspace{1cm} (5)

According to Lemma 1, the cognitive complexity of texts can be reduced to the analysis of the semantic complexity of elements in a given text and the syntactical complexity for composing them into the structure of the sentence and those beyond. Therefore, both the semantic and syntactic complexities of texts are rigorously described in the following before the entire cognitive complexity is formally explained.

**Definition 5.** The semantic complexity of a sentence, $C_{\Theta}$, is a sum of the cognitive weights, $w_i$, $1 \leq i \leq n$, of each of its $n$ words when they are deduced into a known concept or behavior in the given linguistic knowledge base, i.e.:

$$C_{\Theta} = \sum_{i=1}^{n} w_i, n = |W|$$

where the unit of the semantic complexity is a single cognitive weight $(\omega)$ that represents an internal semantic reduction of a terminal word.

**Definition 6.** The cognitive weight of semantic reduction, $w$, in semantic analysis is the comprehension cost for searching or mapping a word in the linguistic knowledge base as shown in Table I.

According to Table I, the deductive search for semantic reduction of a word can be divided into two categories known as the internal and external search. The former is a search in a reader’s internal knowledge base, which results in a unit cost, $w = 1$, when the search and comprehension of the word is successful. The latter is a search in external knowledge bases such as an on-line system, a dictionary, a library, an expert, and a comprehensive study with increasing costs in each type of search as given in Table I. A possible negative outcome of semantic analysis is an exhaustive search that results in an incomprehension with $w = \infty$.

As modeled by the cognitive weights for text comprehension, a reader or a cognitive system first searches the internal knowledge base in order to minimize the semantic analysis cost. However, if any internal search is unsatisfied, an external search must be conducted with higher cost. The cognitive weights also indicate that the subjective properties of individual readers towards the same text can be quantitatively recognized and represented.

**Definition 7.** The syntactic complexity of a sentence, $C_{\Psi}$, is the cognitive cost for composing the individual semantics of words in the sentence into a comprehended sentence.

<table>
<thead>
<tr>
<th>No.</th>
<th>Category</th>
<th>Scope of search</th>
<th>Cognitive weight $(w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal</td>
<td>The reader’s knowledge base</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>External</td>
<td>On-line</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Dictionary</td>
<td>Dictionary</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>Library</td>
<td>Library</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>Expert</td>
<td>Expert</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>A comprehensive study</td>
<td>A comprehensive study</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>Exhaustive</td>
<td>Incomprehension</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

**Table I.** The cognitive weight $(w)$ for semantic reduction.
The Cognitive Complexity of Texts and Its Formal Measurement in Cognitive Linguistics

Wang et al.

The semantic relation denoted by the syntactic rules of the sentence, which is classified into five categories as follows:

\[ C_S = \begin{cases} 
1, S_1 \hat{=} (V \mid \sigma) & \text{meta-sentence} \\
2, S_2 \hat{=} (\text{J} V) & \text{simple sentence} \\
3, S_3 \hat{=} (\text{J} V\emptyset) & \text{complete sentence} \\
4, S_4 \hat{=} (\text{J} PV P NP) & \text{complex sentence} \quad (7) \\
5, S_5 \hat{=} (S\gamma S) & \text{complex sentence with phrases} \\
\end{cases} \]

where the unit of the syntactic complexity is a single composition \([P]\) of the semantics of a single element in a meta-sentence that consists only of a verb or an interjection.

The types of sentences as described in Definition 7 can be illustrated in Figure 2 where the structural influence on semantic composition of sentence elements is demonstrated. According to statistics, that most natural and practical sentences composed in the generic patterns as given in Figure 2 fall into categories 4 or 5, i.e., the complex sentence structures with phrases and clauses in order to convey precise expressions.

According to Lemma 1 as well as Definitions 5 through 7, a quantitative measure of the cognitive complexity of texts can be derived as follows.

**Definition 8.** The cognitive complexity of text, \(C\), is a product of its syntactic complexity \(C_S[P]\) and semantic complexity \(C_\omega[\omega]\), i.e.:

\[ C \hat{=} f(C_S, C_\omega) \]

\[ = C_S \cdot C_\omega \]

\[ = C_S \cdot \sum_{i=1}^{[W]} w_i \quad [P\omega] \quad (8) \]

where the unit of the cognitive complexity is called a composed cognitive weight \((P\omega)\).

**Definition 9.** The physical meaning of a unit of composed cognitive weight, \(P\omega\), denotes a single semantic weight \((\omega)\) by a meta-composition \((P)\) with the simplest structure of a meta-sentence, i.e.:

\[ 1P\omega \hat{=} 1P \cdot 1\omega \quad (9) \]

On the basis of the generic model of the cognitive complexity of texts as developed in Definitions 8 and 9, the cognitive complexities at different levels of the syntactic structures such as those of sentence, paragraph, and essay can be rigorously analyzed from the bottom up in the following section.

**4. MATHEMATICAL MODELS OF THE COGNITIVE COMPLEXITY OF TEXTS**

The cognitive complexity of text comprehension can be modeled and measured at the sentence, paragraph, and essay levels from the bottom up. Then, the metrics of text cognitive complexity is rigorously described by a set of mathematical models in the following subsections.

**4.1. Cognitive Complexity of Texts at the Sentence Level**

**Definition 10.** The comprehension level of texts, \(k\), \(k \in [0,1]\), is determined by the extents of individual reader’s ability to understand the general materials \(k_g\) and professional materials \(k_p\) in the given text as classified in Table II, i.e.:

\[ k = \frac{1}{2}(k_g + k_p), \quad k \in [0,1] \quad (10) \]

The typical comprehensive levels of readers can be classified into four categories known as insufficient \((I = 20\%)\), basic \((B = 60\%)\), medium \((M = 80\%)\), and advanced \((A = 100\%)\). The combinations of these comprehensive levels \((k)\) indicate an individual reader’s average comprehension ability for a given text.

**Definition 11.** The cognitive complexity of a sentence, \(C(S)\), is a product of the syntactical complexity and the semantic complexity of the sentence \((S)\) derived from Eq. (8), i.e.:

\[ C(S) \hat{=} C_S \cdot C_\omega \]

\[ = C_S \cdot \sum_{i=1}^{[W]} w_i \]

\[ = C_S \cdot (\sum_{i=1}^{[W]} w_i + \sum_{j=1}^{(1-\|W\|)} w_j') \]

\[ w_i = 1 \quad \text{(internal)}, \quad 2 \leq w_j' \leq \infty \quad \text{(external)} \]

\[ = C_S \cdot (k\|W\| + \sum_{j=1}^{(1-\|W\|)} w_j') \cdot [P\omega] \quad (11) \]

where \(k, k \in [0,1]\), is the comprehension level of a given text to a reader as determined by Eq. (10), which divides the semantic analyses into \(k\|W\|\) internal searches and \((1-k)\|W\|\) external searches.

The model of sentence complexity as given in Definition 11 explains how the level of comprehension capability of a specific reader is quantitatively measured, which can represent the individual differences on the same sentence due to various comprehension ability and level of knowledge base.

**Corollary 1.** When the average cognitive weight of external search difficulty is a constant, i.e., \(w_i = n, 2 \leq n \leq \infty\), the sentence cognitive complexity is reduced to the follows:

\[ C(S) \hat{=} C_S \cdot C_\omega \]

Corollary 1 indicates that the higher the comprehension level \( k \), the lower the cognitive complexity \( C \) in a text; and vice versa.

**Theorem 2.** The order of the cognitive complexity of texts, \( O(C) \), is constrained by the following limits:

\[
|W| \leq O(C) \leq \infty
\]  

where the lower limit represents a full comprehension, while the upper limit represents an incomprehension after exhausted external searches for semantic reduction.

**Table II.** The comprehension level \( (k) \) of reader’s knowledge bases (KBS).

<table>
<thead>
<tr>
<th>No.</th>
<th>Level of general KB ( (k_g) )</th>
<th>Level of professional KB ( (k_p) )</th>
<th>( k = \frac{1}{2}(k_g + k_p) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>I</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>B</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>I</td>
<td>M</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>I</td>
<td>A</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>I</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>B</td>
<td>0.6</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>M</td>
<td>0.7</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>A</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>I</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>B</td>
<td>0.7</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>M</td>
<td>0.8</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
<td>A</td>
<td>0.9</td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>I</td>
<td>0.6</td>
</tr>
<tr>
<td>14</td>
<td>A</td>
<td>B</td>
<td>0.8</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>M</td>
<td>0.9</td>
</tr>
<tr>
<td>16</td>
<td>A</td>
<td>A</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Weights insufficient: I = 20%; basic: B = 60%; medium: M = 80%; advanced: A = 100%.

**Proof.** Theorem 2 can be directly proven by using Eq. (12), i.e.,:

\[
\begin{align*}
O_{\text{min}}(C) &= \lim_{k \to 1} \left( \frac{1}{2}(n + (1 - n)k)|W| \right) = |W| \\
O_{\text{max}}(C) &= \lim_{k \to 0} \left( \frac{1}{2}(n + (1 - n)k)|W| \right) = n|W| \to \infty
\end{align*}
\]

The cognitive models of text comprehension and its complexity are applicable to both human and machine information processing systems such as intelligent search engines, word and knowledge processors, and cognitive robots.

**4.2. Cognitive Complexity of Texts at the Paragraph Level**

The semantic relations of sentences \( \bullet \) are a set of connectors, which formally model the operations on phrase and sentence compositions and their joint meaning in structures beyond sentences.

**Definition 12.** A paragraph, \( \Psi \), in language \( \mathcal{L} \) is a composition of multiple sentences \( S \) towards a coherent theme of expressions, a behavioral process, and/or a synthesis of a structure, i.e.,:

\[
\Psi \triangleq S \bullet S \quad | S \gamma S \quad | (S_i)^+ \quad | (S_i)^\downarrow (15)
\]

where \( \bullet \) is a set of primitive structural relations between sentences, \( \gamma \subseteq \mathcal{R} \subseteq \mathcal{L} \); \( \gamma \) is a set of coordinatives and correlatives; and \((S_i)^+\) and \((S_i)^\downarrow\) denote that sequential sentences \( \mathcal{S} \) separated by “,” or “;”, respectively, are repeated at least once.

The primary compositional operations of sentences in a paragraph \( \Psi \), \( \bullet \), can be classified as the sequential, parallel, embedded structures, and their combinations known...
as the hybrid structures. The syntactic structures, descriptions, and mathematical models of the primitive sentence structures are summarized in Table III.

In addition to the structural relations of sentences as given in Table III, there are functional relations of sentences such as those of causation (such as \( \therefore \)), extension (such as those of causation (such as \( \therefore \)), explanation (such as \( \therefore \)), therefore, thus, further, moreover, furthermore, ...). However, from a structural point of view, the functional relations are sequential composition of sentences.

**Definition 13.** The *cognitive complexity* of a paragraph \( \Psi \) with the sequential structure of sentences, \( C_s(\Psi) \), is a sum of the cognitive complexities of all individual sentences, i.e.:

\[
C_s(\Psi) \triangleq \sum_{i=1}^{m} C(S_i), \ m = |\Psi|
\]

\[
= \sum_{i=1}^{m} \sum_{j=1}^{s_i} C(S_j) \ [P\omega] \ 
\]

**Definition 14.** The *cognitive complexity* of a paragraph \( \Psi \) with the parallel structure of sentences, \( C_p(\Psi) \), is a sum of the cognitive complexities of all individual sentences multiplied by the number of sentences, \( n \), in the parallel structure, i.e.:

\[
C_p(\Psi) \triangleq n \sum_{i=1}^{n} C(S_i) \ [P\omega] \ 
\]

where \( n \) represents the synchronization cost in parallel semantic analysis.

**Definition 15.** The *cognitive complexity* of a paragraph \( \Psi \) with the embedded structure of sentences, \( C_e(\Psi) \), is a sum of the cognitive complexities of all embedded sentences multiplied by that of the outer layer sentence(s), i.e.:

\[
C_e(\Psi) \triangleq C(S_0) \cdot \sum_{i=1}^{n_c} C(S_i) \ [P\omega] \ 
\]  

(18)

A paragraph with a hybrid structure by combination of the three primitive structures can be treated as a sum of its separated parts that are decomposed to one of the primitive structures.

**Definition 16.** The *cognitive complexity* of a paragraph \( \Psi \) with the hybrid structure of sentences, \( C_h(\Psi) \), is the sum of the cognitive complexities in the three primitive forms of compositional structures, i.e.:

\[
C_h(\Psi) \triangleq C_s(\Psi) + C_p(\Psi) + C_e(\Psi)
\]

\[
= \sum_{i=1}^{n_e} C(S_i) + n_p \sum_{j=1}^{n_c} C(S_j) + C(S_0) \cdot \sum_{i=1}^{n_e} C(S_i) \ [P\omega] \ 
\]

(19)

### 4.3. Cognitive Complexity of Texts at the Essay Level

The structure and compositional relations of paragraphs in an essay (\( \Psi \)) can be treated as a sequential structure among all paragraphs. Therefore, the measurement of the cognitive complexity of essays can be derived as the sum of all its paragraphs.
**Definition 17.** The cognitive complexity of an essay $\mathcal{E}$ with sequential paragraph structures, $C(\mathcal{E})$, is the sum of the cognitive complexity of the $p$ paragraphs, i.e.:

$$C(\mathcal{E}) \geq \sum_{i=1}^{p} C(\Psi_i), p = |\mathcal{E}|$$

$$= \sum_{i=1}^{p} C(\Psi_i) \cdot |P\omega|$$

**Definition 18.** The average cognitive complexity per word in an essay or lower structures, $r(\mathcal{E})$, is a ratio between the sum of the cognitive complexity of the text $C(\mathcal{E})$ and the total number of words in the text $|W|$, i.e.:

$$r(\mathcal{E}) = \frac{C(\mathcal{E})}{|W|} \cdot |P\omega/W|$$

The measurement framework of the cognitive complexity of text systematically developed at the sentence, paragraph, and essay levels provide a coherent metrics for quantitative complexity analysis of texts in natural languages.

**5. Applications of the Text Cognitive Complexity Metrics in Cognitive Linguistics**

The measurement framework of the cognitive complexity of texts has been developed in preceding sections. This section presents a set of case studies on the measurement of text comprehension complexities to illustrate the applications of the theory and methodology for text complexity analyses in English as a typical natural language.

**Example 1.** For a sentence $S_1 = \text{"Claud, who was groaning and growling, got up and limped off."}$. Assuming both general and professional comprehensibility of a reader $R_1$ are advanced, i.e., $k = 1.0$, analyze the cognitive complexity of $R_1$ for the comprehension of $S_1$.

According to Eq. (8), as an advanced reader with $k = 1$ in text comprehension, only internal searches are needed. Therefore, the cognitive complexity for the given sentence $C(S_1 | R_1)$ is:

$$C(S_1 | R_1) = C_\psi \cdot C_\omega$$

$$= C_\psi \cdot \sum_{i=1}^{|W|} w_i$$

$$= C_\psi \cdot |W|$$

$$= 5 \cdot 11.0$$

$$= 55.0 [P\omega]$$

Use the same experiment for another reader $R_2$, with an insufficient comprehensibility, e.g., $k = 0.3$. Assuming the average cognitive weight for external search is $\overline{w} = 3$, i.e., a combined use of dictionary and on-line knowledge bases for the $1 - k = 70\%$ external searches, the cognitive complexity for comprehending $S_1$ with regards to $R_2$ is as follows:

The above examples demonstrate that the cognitive complexity of text provides a quantitative explanation for the subjective complexity of text in natural languages for different individuals. It is observed that there is a certain empirical cognitive threshold ($\tau$) for sentence comprehension where $\tau = 100 \cdot P\omega$. Beyond this threshold, the comprehension difficulty of the given text will be significantly high for a specific reader. The cognitive threshold can be refined as the average cognitive complexity per word in the following corollary.

**Corollary 2.** The average cognitive complexity per unit word in a given text at the levels of sentence $r(S)$, paragraph $r(\Psi)$, and essay $r(\mathcal{E})$ is a ratio between the cognitive complexity $C$ and the number of words in the text $|W|$ at corresponding structural levels, i.e.:

$$r(S) \geq \frac{C(S)}{|W_S|} \cdot |P\omega/W|$$

$$r(\Psi) \geq \frac{C(\Psi)}{|W_\Psi|} \cdot |P\omega/W|$$

$$r(\mathcal{E}) \geq \frac{C(\mathcal{E})}{|W_\mathcal{E}|} \cdot |P\omega/W|$$

where the threshold for a complex text is $r \geq 10.0 \cdot P\omega/W$.

**Example 2.** Take $S_1$ for $R_2$ in Example 1 as an instance, the average cognitive complexity per unit word, $r(S_1 | R_2)$, can be quantitatively estimated as:

$$r(S_1 | R_2) = \frac{C(S_1 | R_2)}{|W_S|}$$

$$= \frac{132.0}{11}$$

$$= 12.0 \cdot P\omega/W$$

The result shows that $S_1$ is a semantically hard sentence to $R_2$ because $r(S_1 | R_2) > r = 10.0 \cdot P\omega/W$. However, $r(S_1 | R_1)$ is significantly lower as analyzed below:

$$r(S_1 | R_1) = \frac{C(S_1 | R_1)}{|W_S|}$$

$$= \frac{55.0}{11}$$

$$= 5.0 \cdot P\omega/W$$

because reader $R_1$ is much advanced who understand the semantics of all words in $S_1$ as given in Example 1.

$$C(S_1 | R_2) = C_\psi \cdot C_\omega$$
Example 3. Given a sentence, \( S_2 = \) “A cognitive computer is an intelligent computer for knowledge processing.” Assuming the distribution of cognitive weights of readers \( R_1 \) and \( R_4 \) are as given in Table IV, analyze the cognitive complexities of \( S_2 \) with regards to readers \( R_1 \) and \( R_4 \), respectively.

Applying Eq. (8), the cognitive complexities of \( S_2 \) with regards to Reader 3 and Reader 4 can be quantitatively measured as follows, respectively:

\[
C(S_2 | R_1) = C_q \cdot C_{\phi} \cdot \frac{|W|}{|S|} = C_q \cdot \sum_{i=1}^{[W]} w_i = 4 \cdot (1 + 6 + 3 + 1 + 1 + 4 + 1 + 1 + 2 + 2) = 88 \ P\omega
\]

and

\[
C(S_2, R_4) = C_q \cdot C_{\phi} \cdot \frac{|W|}{|S|} = C_q \cdot \sum_{i=1}^{[W]} w_i = 4 \cdot 10 = 40 \ P\omega
\]

It is obvious that the cognitive complexity of sentence \( S_2 \) is much higher to reader \( R_1 \) than \( R_4 \) because the latter only needs internal searches in semantic reduction.

Example 4. Analyze the cognitive complexity of the following sample paragraph \( \Psi_1 \) with complex sentences for Readers \( R_k \) (\( k = w = 1 \)) and \( R_w \) (\( k = 0 \) and \( w = 2.5 \)) as follows:

“The general principles of dynamics are rules which demonstrate a relationship between the motions of bodies and the forces which produced those motions. Based in large part on the work of his predecessors, Isaac Newton deduced three laws of dynamics which he published in 1687 in his famous Principia. These laws are collectively known as Newton’s laws of motion, which provide the basis for understanding the effect that forces have on an object.”

This paragraph encompasses three sentences in a sequential structure. All sentences are complex sentence with clauses whose syntactic complexities are 5. According to Eqs. (8) and (22), the cognitive complexity of \( \Psi_1 \) with regards to reader \( R_3 \) is:

\[
C(\Psi_1 | R_3) = \sum_{i=1}^{3} C(S_i)
= \sum_{i=1}^{3} (C_q \cdot \sum_{j=1}^{[W]} w_j)
= 5 \cdot (23 + 26 + 24)
= 365 \ P\omega
\]

\[
r(\Psi_1 | R_3) = \frac{C(\Psi_1 | R_3)}{|W|} = \frac{365}{73} = 5.0 \ SP/W
\]

For another reader \( R_6 \) whose average cognitive weight for words in \( \Psi_1 \) is assumed as \( \bar{w} = 2.5 \), the cognitive complexity on the same paragraph, \( C(\Psi_1 | R_6) \), is as follows:

\[
C(\Psi_1 | R_6) = \sum_{i=1}^{3} C(S_i)
= \sum_{i=1}^{3} (C_q \cdot \sum_{j=1}^{[W]} w_j)
= 5 \cdot (2.5 \cdot (23 + 26 + 24))
= 912.50 \ P\omega
\]

\[
r(\Psi_1 | R_6) = \frac{C(\Psi_1 | R_6)}{|W|} = \frac{912.50}{73} = 12.50 \ P\omega/W
\]

The analysis results indicates that \( C(\Psi_1 | R_6) = 5.0 \ P\omega/W \gg C(\Psi_1 | R_3) = 12.5 \ P\omega/W \), because of the extensive needs of external searches for semantic reduction.

Example 5. Analyze the cognitive complexity of the following sample paragraph \( \Psi_2 \) for a reader \( R_7 \) with \( w = 1 \):

“Many commentators have noted that sentence connectors are an important and useful element in expository and

<table>
<thead>
<tr>
<th>Word (Reader 3)</th>
<th>A</th>
<th>cognitive computer is an intelligent computer for knowledge processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_j )</td>
<td>1</td>
<td>6 3 1 1 4 1 1 2 2</td>
</tr>
<tr>
<td>( w_j )</td>
<td>1</td>
<td>1 1 1 1 1 1 1 1 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word (Reader 4)</th>
<th>A</th>
<th>cognitive computer is an intelligent computer for knowledge processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_j )</td>
<td>1</td>
<td>6 3 1 1 4 1 1 2 2</td>
</tr>
<tr>
<td>( w_j )</td>
<td>1</td>
<td>1 1 1 1 1 1 1 1 1</td>
</tr>
</tbody>
</table>
argumentative writing. Frequency studies of their occurrence in academic English extended at least as far back as Huddleston (1971). Textbooks have for many years regularly included chapters on sentence connectors. Most reference grammars deal only with their grammatical status, classification, meaning, and use.”

Note that the paragraph $\Psi_2$ is a series structure of sentences. According to Eqs. (8) and (22), the cognitive complexity of $\Psi_2$ with regards to reader $R_1$ is:

$$C(\Psi_2 | R_1) = \sum_{i=1}^{4} C(S_i)$$

$$= \sum_{i=1}^{4} (C_{\Psi} \cdot \sum_{j=1}^{W_i} w_j)$$

$$= 5 \cdot 18 + 4 \cdot 17 + 4 \cdot 11 + 4 \cdot 13$$

$$= 254 \text{ Pwo}$$

$$r(\Psi_2 | R_1) = \frac{C(\Psi_2 | R_1)}{|W|} = \frac{254}{59} = 4.31 \text{ Pwo/W}$$

The case studies presented in this section reveal that the cognitive complexity of texts not only provides a precise measurement for text comprehension, but also enables a rigorous explanation for the mechanisms of the cognitive complexity of text in natural languages. A wide range of applications in cognitive linguistics and computational linguistics can be derived based on the basic study on the fundamental properties of natural languages and their rigorous measurement.

6. CONCLUSIONS

An important and conventionally unquantifiable problem on the cognitive complexity of texts in cognitive linguistics has been rigorously modeled by using denotational mathematics. The cognitive foundations of text comprehension and the discourse of formal natural languages have been systematically explored. The mathematical models of cognitive complexity of texts have been created as the product of the syntactic and semantic complexities of texts at the sentence, paragraph, and essay levels. A set of case studies have been presented to demonstrate the applications of this work in web text processing, search engines, natural language processing, computational linguistics, cognitive computing, cognitive learning engines, and computing with words.

This work has revealed that the cognitive complexity of texts is proportional to both the syntactic complexity and the semantic complexity. It has also found that the threshold of the average cognitive complexity per unit word in a given text is $r = 10 \text{ Pwo/W}$, beyond this threshold a reader will feel empirical difficulty for literature comprehension. Conventionally, the syntax and semantics of languages were treated separately, where the syntax has been statically perceived as only for structural analysis of a sentence. However, this work has revealed that the syntax of a sentence in natural languages plays an important role in both structural and semantic analyses where syntax guides the syntheses of semantics at the word and phrase levels as composition rules. This finding has explained that the cognitive process of text comprehension is a closed circle of both analysis and synthesis where:

(a) The syntactic analysis breaks up a sentence into its elements in order to enabling the semantic analysis of words;

(b) The composition of the semantics of words in the sentence based on the syntactic rules results in the comprehension of meaning of the entire sentence.

Therefore, the syntax of a language not only provides rules for semantic analyses in the process of semantic reduction from the top done, but also provides rules for semantic synthesis in order to reassemble the composed semantics of the entire sentence from the bottom up.

Acknowledgments: The authors would like to acknowledge the Natural Science and Engineering Council of Canada (NSERC) for its partial support to this work. We would like to thank the anonymous reviewers for their valuable comments and suggestions.

References


The Cognitive Complexity of Texts and Its Formal Measurement in Cognitive Linguistics

Wang et al.


Received: 8 July 2012. Accepted: 1 November 2012.