Towards a Formal Framework of Cognitive Linguistics

Yingxu Wang¹ and Robert C. Berwick ²

¹ International Institute of Cognitive Informatics and Cognitive Computing (IICICC)
Laboratory for Cognitive Informatics and Cognitive Computing
Dept. of Electrical and Computer Engineering, Schulich School of Engineering
University of Calgary
2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4
Email: yingxu@ucalgary.ca

² Dept. of Brain and Cognitive Science and Dept. Electrical Engineering and Computer Science
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139, USA
Email: berwick@csail.mit.edu

Abstract: Cognitive linguistics is an emerging discipline that studies the cognitive properties of natural languages and the cognitive models of languages in computing linguistics, cognitive computing, and computational intelligence. This paper presents the theoretical framework of cognitive linguistics in order to systematically formalize the syntaxes and grammars of natural languages. An abstract language model of cognitive linguistics is created at the top level. Based on it, the cognitive structures of languages at the levels of lexis, phrase, clauses, sentence, paragraph, and article are formally modeled from the bottom up. Using contemporary denotational mathematics, the deductive grammar of English is formally modeled and rigorously analyzed. This basic research provides support for a wide range of applications in computational linguistics, cognitive informatics, online text processing, web search engines, machine language comprehension, autonomous machine learning, smart cell phones, semantic computing, and computing with words.

Keywords: Cognitive informatics, cognitive computing, cognitive linguistics, natural language processing, lexical models, sentence models, paragraph models, article models, deductive grammar of English, deductive syntax, deductive semantics, concept algebra

1. Introduction

It is recognized that the development of formal models of syntaxes, semantics, and grammars of natural languages is a fundamental challenge in cognitive linguistics. The mathematical models of natural languages are not only required in cognitive informatics, natural language processing, and computational linguistics in general, but also needed in word processing, web search engines, machine language comprehension, autonomous machine learning, smart cell phones, semantic computing, cognitive computing, and computing with words in particular. Without the mathematical models of grammars and linguistics, the empirical language knowledge of humans cannot be systematically and precisely conveyed to cognitive machines and systems, because the syntaxes and semantics of languages might not be exhaustedly enumerated by examples.

**Definition 1.** Cognitive linguistics is an emerging discipline that studies the cognitive properties of natural languages and the cognitive models of languages in cognitive computing and computational intelligence.

In order to rigorously explain the cognitive mechanisms of languages, syntaxes and semantics of a given language as well as their relations and interactions have to be systematically studied. In a language, syntaxes deal with relations and combinational rules of words in sentence formation; while semantics embody meanings of words and sentences. Syntactic and semantic analyses in linguistics rely on a set of explicitly described rules known as the grammar of the language. Therefore, contemporary linguistic analyses focus on the study of grammars, which is centered in language acquisition, understanding, and interpretation.

**Definition 2.** The grammar of a language is a set of rigorously specified rules that integrates phonetics, phonology, morphology, syntax, and semantics of the language.

Grammars govern the articulation, perception, and patterning of speech sounds, the formation of phrases and sentences, and the interpretation of utterance. However, the formalization of grammars by mathematical models is not only a fundamental need, but also a significant challenge in cognitive and computational linguistics.

This paper attempts to demonstrate that natural languages can be systematically formalized by using denotational mathematics and cognitive informatics methodologies. In the remainder of the paper, Section 2 formally describes the discourse of natural languages and its mathematical model at the top level. Section 3 analyzes and formally describes the discourse of natural languages and its denotational mathematics and cognitive informatics. Languages can be systematically formalized by using only a fundamental need, but also a significant challenge in formalization of grammars by mathematical models is not sentential and the interpretation of utterance. However, the formalization of grammars by mathematical models is not only a fundamental need, but also a significant challenge in cognitive and computational linguistics.

**2. The General Mathematical Model of Natural Languages**

A formal model of a general language such as English can be described as an abstract language using denotational mathematics [Wang, 2008a, 2011]. Based on the general abstract language model, the structural models of natural languages can be created at the lexis, phrase, clause, sentence, paragraph, and article levels from the bottom up.

**Definition 3.** The abstract language, $L$, is a 5-tuple, i.e.:

\[
L = (\mathcal{E}, \mathcal{X}, \mathcal{W}, \mathcal{R}, \mathcal{S})
\]  

where

- $\mathcal{E}$ is a finite ordered set of alphabet of the language, $\mathcal{E} \subseteq L$;
- $\mathcal{X}$ is a power set of lexical relations between the letters in the alphabet, $\mathcal{X} \subseteq \mathcal{E} \subseteq L$;
- $\mathcal{W}$ is a finite nonempty set of words that are identified strings of the alphabet in the language, $\mathcal{W} = \{ \{ \mathcal{E} \mid \mathcal{X} \} \subseteq L \}$. A subset of $\mathcal{W}$ is known as the primitive words ($\mathcal{W}^p$), $\mathcal{W}^p \subseteq \mathcal{W}$, that directly represents real-world entities, proper names, meta-behaviors, and abstract concepts that cannot be further deduced onto more primitive concepts or behaviors;
- $\mathcal{R}$ is a power set of syntactic relations between words, $\mathcal{R} = \mathcal{W} \times \mathcal{W} \subseteq L$;
- $\mathcal{S}$ is a power set of semantic relations between words, $\mathcal{S} = \mathcal{R} \rightarrow \mathcal{W}^p = \mathcal{W} \times \mathcal{W} \rightarrow \mathcal{W}^p \times \mathcal{W}^p$, $\mathcal{R} \subseteq \mathcal{R} \subseteq L$.

The formal grammar for a language can be categorized as the syntactical and semantic grammars. The former are grammars specifying the structural rules of the language; while the latter are grammars specifying the functional rules of the language. A special part of the semantic grammar of natural languages is the set of words, or the vocabulary, which can be classified as terminals and nonterminals.

**Definition 4.** The vocabulary of a language, $\mathcal{V}$, is a finite nonempty set of words that represent certain meanings or convey specific semantics, i.e.:

\[
\mathcal{V} = \mathcal{T} \cup \overline{\mathcal{T}} \subseteq \mathcal{W} \subseteq L
\]  

where $\mathcal{T}$ and $\overline{\mathcal{T}}$ denote the sets of terminals and nonterminals, respectively, in the languages.

**Definition 5.** The terminals of a language, $\overline{\mathcal{T}}$, is a finite nonempty subset of $\mathcal{V}$ in which the vocabulary fall into the following special primitive categories, i.e.:

\[
\mathcal{T} \triangleq \{ \mathcal{V}_0 \mid \mathcal{V}_0 \subseteq (N_{\text{e0}} \cup N_{\text{a0}} \cup V_0 \cup M_0) \subset \mathcal{W} \}
\]  

where $N_{\text{e0}}$, $N_{\text{a0}}$, $V_0$, and $M_0$ denote words that represent real-world entities, meta-abstract concepts, meta-behavioral verbs, and common modifiers, respectively.

**Definition 6.** The nonterminals of a language, $\overline{\mathcal{T}}$, is a subset of $\mathcal{V}$ in which the vocabulary can be defined or derived by the terminals, i.e.:

\[
\overline{\mathcal{T}} = \mathcal{V} \setminus \mathcal{T}
\]
The abstract language $L$ as given in Definition 3 explains the universal discourse of languages as the contexts of linguistics. On the basis of the abstract language model $L$, the fundamental structural properties of natural languages at the lexis, phrase, clause, sentence, paragraph, and article levels can be rigorously modeled and analyzed in the following sections.

### 3. The Formal Lexical Structures of English

A set of three categories and 19 lexical and syntactic elements in language $L$ is elicited and summarized in Table 1. The lexes are a category of syntactic units that represent the six basic parts of speech. The modifiers are a category of syntactic units that refine and restrict the semantics of words. The phrases are a category of syntactic units that compose multiple words into a complex unit.

**Definition 7.** The set of lexical elements $E$ of a sentence $S$ in language $L$ can be classified into the categories of lexes ($\mathcal{L}$), modifiers ($\mathcal{M}$), and phrases ($\mathcal{P}$), i.e.:

$$E \triangleq (\mathcal{L}, \mathcal{M}, \mathcal{P})$$

$$= \{N, X, V, A, D, P\}$$

$$\ | \ \{\tau, \kappa, \delta, \alpha, \neg, \gamma, \sigma\}$$

$$\ | \ \{NP, VP, AP, DP, PP, CP\}$$

Each of the lexical elements $E$ in English is formally modeled and explained in the following subsections.

#### 3.1 Parts of Speech

**Definition 8.** The lexes, $L$, in language $L$ encompass six categories of lexical roles known as the parts of speech such as noun, pronoun, verb, adjective, adverb, and preposition where each category can be formally modeled as a specific set of words in the lexical categories $N$, $X$, $V$, $A$, $D$, and $P$, respectively, in $L$, i.e.:

$$N \triangleq \{ n \in N \mid N \subset W \subset L \}$$

$$X \triangleq \{ x \in X \mid X \subset W \subset L \}$$

$$V \triangleq \{ v \in V \mid V \subset W \subset L \}$$

$$A \triangleq \{ a \in A \mid A \subset W \subset L \}$$

$$D \triangleq \{ d \in D \mid D \subset W \subset L \}$$

$$P \triangleq \{ p \in P \mid P \subset W \subset L \}$$

### Table 1. The Syntactic Categories of Linguistics

<table>
<thead>
<tr>
<th>No.</th>
<th>Category</th>
<th>Subcategory</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lexis</td>
<td>$\mathcal{L}$</td>
<td>Basic parts of speech</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Noun</td>
<td>$N$</td>
<td>Entities, abstract concepts, and behaviors</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>Pronoun</td>
<td>$X$</td>
<td>Representation of nouns</td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td>Verb</td>
<td>$V$</td>
<td>Actions, states, and possessions</td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>Adjective</td>
<td>$A$</td>
<td>Properties of a noun</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>Adverb</td>
<td>$\Delta$</td>
<td>Properties of a verb</td>
<td></td>
</tr>
<tr>
<td>1.6</td>
<td>Preposition</td>
<td>$P$</td>
<td>Relations between two words or phrases</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Modifier</td>
<td>$\mathcal{M}$</td>
<td>Semantic refinement and restrictive</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Determiner</td>
<td>$\tau$</td>
<td>Articles and special pronouns or adjectives that introduce or restrict a noun</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>Qualifier</td>
<td>$\kappa$</td>
<td>Special adverbs that modify or restrict the extent or manner of a verb</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>Degree</td>
<td>$\delta$</td>
<td>Special adverbs or adjectives that modify or restrict the extent of an adjective or an adverb</td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td>Auxiliary</td>
<td>$\alpha$</td>
<td>Verbal prefixes that specifies the tense, voice, mood, person, and/or number of a verb</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>Negative</td>
<td>$\neg$</td>
<td>Special words that express the opposite meaning of a phrase or sentence</td>
<td></td>
</tr>
<tr>
<td>2.6</td>
<td>Conjunction</td>
<td>$\gamma$</td>
<td>Connections between equivalent words, phrases, and clauses such as coordinating, subordinating, and correlatives</td>
<td></td>
</tr>
<tr>
<td>2.7</td>
<td>Interjection</td>
<td>$\sigma$</td>
<td>A word of emotional expression</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Phrase</td>
<td>$\mathcal{P}$</td>
<td>A syntactic unit with multiple words</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Noun phrase</td>
<td>NP</td>
<td>A composition of a noun with a determiner or adjective (phrase)</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Verb phrase</td>
<td>VP</td>
<td>A composition of a verb with a noun (phrase) or adverb (phrase)</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Adjective phrase</td>
<td>AP</td>
<td>A composition of an adjective with a degree word or noun (phrase)</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Adverb phrase</td>
<td>DP</td>
<td>A composition of an adverb with a determiner or verb (phrase)</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>Prepositional phrase</td>
<td>PP</td>
<td>A composition of a preposition with a determiner or noun (phrase)</td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td>Complement phrase</td>
<td>CP</td>
<td>A composition of an AP, NP, or DP to a subject ($\mathcal{S}$) or object ($\mathcal{O}$), and a predicate (VP).</td>
<td></td>
</tr>
</tbody>
</table>
An important property of lexes is known as inflections, which represents the morphological flexibility of natural languages that allows the change of forms of words in accordance with grammatical roles, relations, and agreements between parts of speech in a sentence. Each part of speech including related inflection rules will be rigorously specified in the following subsections with formal models.

**Definition 9.** A noun, \( N \), in language \( \mathcal{L} \) is a part of speech that identifies and represents a real-world entity, a proper name, a meta-behavior, or an abstract concept in the following pattern:

\[
N \triangleq N_0 | N_1 | N_n \quad / / \text{Mass, singular, and plural nouns}
\]

\[
| N'_o \quad / / \text{Irregular plural nouns}
\]

\[
| N'_p \quad / / \text{Proper nouns}
\]

\[
| X \quad / / \text{Pronouns}
\]

\[
| V_{\text{vog}} \quad / / \text{Gerunds}
\]

\[
| N^+ \quad / / \text{Complex nouns}
\]

\[
| (N_i)^+ \quad / / \text{Co-occurrence}
\]

\[
| N \gamma N
\]

where \( \gamma \) represents a conjunction that will be defined in the category of modifiers, and \( N^+ \) represents a compound noun that repeat \( N \) for one or more times such as \( NN \) and \( N, N, \ldots, N \), respectively.

In addition to the lexical and syntactical patterns, a set of relational grammar rules between nouns and verbs, pronouns, and determiners can be specified known as \( R_2 \) (number), \( R_3 \) (tense), \( R_4 \) (mood), \( R_5 \) (number), \( R_6 \) (gender), \( R_7 \) (role), and \( R_8 \) (determiner), respectively. Because of the limit of space, the mathematical models of the relational grammar rules will be reported separately [Wang & Berwick, 2012].

**Definition 10.** A pronoun, \( X \), in language \( \mathcal{L} \) is a part of speech that is used to represent or indicate a personal, impersonal, and proper noun in nine categories in the following pattern:

\[
X \triangleq X_{01} | X_{0n} \quad / / \text{Impersonal, single/plural}
\]

\[
| X_{11} | X_{1n} \quad / / \text{First person, single/plural}
\]

\[
| X_{21} | X_{2n} \quad / / \text{Second person, single/plural}
\]

\[
| X_{31m} | X_{3nm} \quad / / \text{Third person (male), single/plural}
\]

\[
| X_{31f} | X_{3nf} \quad / / \text{Third person (female), single/plural}
\]

\[
| X_{p} \quad / / \text{General person}
\]

\[
| X_{il} | X_{in} \quad / / \text{Indefinite, single/plural}
\]

\[
| X_{01l} | X_{0nl} \quad / / \text{Demonstrative, single/plural}
\]

\[
| X_{wh} \quad / / \text{Interrogative}
\]

where the first six categories are personal or impersonal pronouns; while the last three categories are relative pronouns for introducing indefinite, demonstrative, and interrogative noun phrase (NP) and/or subordinate clauses.

It is noteworthy that most pronouns except the relative pronouns have inflections in the forms of subjective \((X_1)\), objective \((X_2)\), possessive \((X_3)\), and possessive-noun \((X_4)\) pronouns as shown in the second part of Eq. 8. A set of relational grammar rules between pronouns and nouns or verbs can be described by \( R_1 \) (person), \( R_3 \) (tense), \( R_4 \) (mood), \( R_5 \) (gender), \( R_7 \) (role), and \( R_8 \) (determiner), respectively [Wang & Berwick, 2012].

**Definition 11.** A verb, \( V \), in language \( \mathcal{L} \) is a part of speech that denotes a state, an existence, a behavior, an action, and an occurrence in the following pattern:

\[
\forall \triangleq V \quad / / \text{Stems}
\]

\[
| V_o \quad / / \text{3rd-person-singular (regular)}
\]

\[
| V'_s \quad / / \text{3rd-person-singular (irregular)}
\]

\[
/ / \text{Relation to objects}
\]

\[
| V_1 \quad / / \text{Transitive}
\]

\[
| V_i \quad / / \text{Intransitive}
\]

\[
| V_{\text{it}} \quad / / \text{Ditransitive}
\]

\[
/ / \text{Tenses}
\]

\[
| V_p = (V | V_1 | V'_s) \quad / / \text{Present}
\]

\[
| V'_f = \alpha V \quad / / \text{Future}
\]

\[
| V_{\text{pt}} \quad / / \text{Past}
\]

\[
| V_{\text{p}} \quad / / \text{Past (irregular)}
\]

\[
| V_{\text{pp}} \quad / / \text{Past participle}
\]

\[
| V_{\text{pp}} \quad / / \text{Past participle (irregular)}
\]

\[
/ / \text{Nonfinite (verbals)}
\]

\[
| V_{\text{to}} \quad / / \text{Infinitives}
\]

\[
| V_{\text{vog}} \quad / / \text{Gerunds}
\]

\[
| V_{\text{vog}} \quad / / \text{Present participial}
\]

\[
| V_{\text{v}} \quad / / \text{Past participial}
\]

\[
/ / \text{Special verbs}
\]

\[
| BE = be | am | is | are | was | were | to be | being | been
\]

\[
| HAVE = have | has | had
\]

\[
| DO = do | does | did | done
\]

\[
/ / \text{Complex verbs}
\]

\[
| (V_1, V'_s)
\]

\[
\forall \gamma \forall
\]

(9)
The inflections of verbs are specified in Eq. 9 according to person, object, tense, and their nonfinite forms. The irregular forms of verbs, particular those of frequently used ones such as BE, HAVE, and DO, are highlighted in Eq. 9.

The grammar rules of verbs encompass those of person, number, tense, mood, and voice, as well as modifiers and special structures. A set of relational grammar rules between verbs and nouns, pronouns, modifiers, and special sentence structures such as R₁ (person), R₂ (number), R₃ (tense), R₄ (mood), R₅ (voice), R₆ (qualifier), R₇ (negation), R₈ (inverted order), and R₉ (There BE structure) is specified, respectively, [Wang & Berwick, 2012].

**Definition 12.** An adjective, A, in language L is a part of speech that is used to modify or restrict a noun or pronoun in the following pattern:

\[ A \triangleq A \] // Positive

\[ | A_{cr} \] // Comparative

\[ | A_{sp} \] // Comparative (irregular)

\[ | A_{sl} \] // Superlative

\[ | A_{sl} \] // Superlative (irregular)

\[ | (N_0's \mid N_1's) \] // Possessive nouns (singular)  \( \text{(10)} \)

\[ | N_{sp} \] // Possessive nouns (plural)

\[ | X_{pp} \] // Possessive pronouns

\[ | V_{toP} \] // Infinitive phrases

\[ | V_{edP} \] // Participial phrases

\[ | PP \] // Propositional phrase

\[ | (A_{s})' \] // Complex adjectives

\[ | A \] // Indefinite

where \( A, A_{cr}, \) and \( A_{sp} \) represent the positive, comparative, and superlative levels of adjectives, respectively.

**Definition 13.** An adverb, D, in language L is a part of speech that is used to modify or restrict a verb, an adjective, or another adverb in the following pattern:

\[ D \triangleq D \] // Positive

\[ | (more \mid much \mid less) D \] // Comparative

\[ | D_{c} \] // Comparative (irregular)

\[ | (most \mid least) D \] // Superlative

\[ | D_{s} \] // Superlative (irregular)

\[ | V_{sp} \] // Infinitive phrases

\[ | PP \] // Propositional phrase

\[ | (D_{s})' \] // Complex adjectives

\[ | D \] // Indefinite

where \( D, D_{cr}, \) and \( D_{sp} \) represent the positive, comparative, and superlative levels of adverbs, respectively.

**Definition 14.** A preposition, P, in language L is a part of speech that is used to link and relate a word phrase, or clause to another counterpart in the following pattern:

\[ P \triangleq P \] // Primitive

\[ | P \gamma P \] // Complex adjectives  \( \text{(12)} \)

**3.2 Modifiers**

The parts of speech described in the preceding subsection describe the roles of words in a language. When analyzing relations of words in a sentence, some words behave as modifiers that affect the syntax or semantics of the target words. The modifiers are a subset of adjectives, adverbs, pronouns, nouns, and infinite verbs.

**Definition 15.** The modifiers in L are words or phrases that describe, limit, qualify or connect another word or phrase in the categories of determiners (τ), qualifiers (κ), degrees (δ), auxiliaries (α), negatives (¬), conjunctions (γ), and interjections (σ) where each category can be formally modeled as a specific set in L, i.e.:

\[ \tau \triangleq \{ \tau_i \mid \tau \sim N \land \tau \subseteq (T \cup X \cup A) \subset W \subset L \} \]

\[ \kappa \triangleq \{ \kappa_i \mid \kappa \sim V \land \kappa \subseteq D \subset W \subset L \} \]

\[ \delta \triangleq \{ \delta_i \mid \delta \sim (A \cup D) \land \delta \subseteq D \subset W \subset L \} \]

\[ \alpha \triangleq \{ \alpha_i \mid \alpha \sim V \land \alpha \subseteq U \subset W \subset L \} \]

\[ \neg \triangleq \{ \neg_i \mid \neg \sim (P \cup S) \land \neg \subseteq (P \cup A \cup X) \subset W \subset L \} \]

\[ \gamma \triangleq \{ \gamma_i \mid \gamma \subset W \subset L \} \]

\[ \sigma \triangleq \{ \sigma_i \mid \sigma \subset W \subset L \} \]

where \( \sim \) represents a relation between a certain type of modifies and the suitable type(s) of elements that it may operate.

Each category of modifiers will be rigorously specified in the following subsections with formal models.

**Definition 16.** The determiners, τ, in language L are a set of articles, demonstrative adjectives, restrictive adjectives, and possessive pronouns that introduces or restricts a noun in the following pattern:

\[ \tau \triangleq ARTICLE \mid Demonstrative_A \mid Restrictive_A \mid Possessive_X \]

\[ = \{ (ARTICLE = \{a, an, the\}) \mid (Demonstrative_A = \{this, that, these, those\}) \mid (Restrictive_A = \{every, each, any, all, some, ...\}) \mid (Possessive_X = \{my, his, her, our, your, their, its\}) \} \]

(14)

where \( \tau \subseteq (T \cup X \cup A) \).

The number rules of determiners are specified by R₉ in [Wang & Berwick, 2012].
Definition 17. A qualifiers, $\kappa$, in language $\mathcal{L}$ are a set of special adverbs that modifies the extent or manner of a verb as follows:

$$\kappa \triangleq \{\text{almost, always, often, perhaps, never, ...}\}, \kappa \subset \mathcal{D}$$  \hspace{1cm} (15)

The modification rules of qualifiers are specified by R$_{10}$ in [Wang & Berwick, 2012].

Definition 18. The degrees, $\delta$, in language $\mathcal{L}$ are a set of special adverbs or adjectives that modifies or restricts the extent of an adjective or an adverb as follows:

$$\delta \triangleq \{\text{too, so, very, more, quite, ...}\}, \delta \subset (\mathcal{A} \cup \mathcal{D})$$  \hspace{1cm} (16)

The modification rules of degrees are specified by R$_{11}$ in [Wang & Berwick, 2012].

Definition 19. The auxiliaries, $\alpha$, in language $\mathcal{L}$ is a set of verbal prefixes that indicates the tense, voice, mood, person, and/or number of a verb in the following pattern:

$$\alpha \triangleq \text{ASPECT} \mid \text{Positive MODAL} \mid \text{Passive MODAL} \mid \text{Durant MODAL}$$

\hspace{1cm} (17)

The tense rules of auxiliaries are specified in R$_{3}$ in [Wang & Berwick, 2012].

Definition 20. The negations, $\neg$, in language $\mathcal{L}$ is a category of words that expresses the opposite meaning of a phrase or sentence, i.e.:

$$\neg \triangleq \{\text{not, no, never, nothing, nobody, none, neither...nor, ...}\}, \neg \subset (\mathcal{D} \cup \mathcal{A} \cup \mathcal{X})$$  \hspace{1cm} (18)

The negation rules are specified by R$_{12}$ in [Wang & Berwick, 2012].

Definition 21. The conjunctions, $\gamma$, in language $\mathcal{L}$ is a special part of speech that connects and relates words, phrases, clauses, by using the coordinatives, subordinates, and correlatives in the following pattern:

$$\gamma \triangleq (\gamma_{\text{Coordinatives}} \mid \gamma_{\text{Subordinates}} \mid \gamma_{\text{Correlatives}})$$

\hspace{1cm} (19)

It is noteworthy in Eq. 19 that the coordinative and correlative conjunctions connect two or more equivalent elements in parallel where the elements can be at the levels of word, phrase, clause, and sentence. However, the subordinate conjunctions relate a subsentence to the main sentence.

Definition 22. The interjections, $\sigma$, in language $\mathcal{L}$ is a special part of speech that expresses an exclamation, qualification, or degree, which extends the context of a sentence in the following pattern:

$$\sigma \triangleq (\text{Exclamation} \mid \text{qualifiers ($\kappa$) \mid degrees ($\delta$)})$$

\hspace{1cm} (20)

The interjections usually appear in the front of sentence separated by the exclamation or comma to denote a strong or weak emotional expression, respectively.

3.3 Phrases

Phrases are the second-level structures beyond lexes in a sentence. A phrase is not a complete sentence where its principal verb is not in the finite form. Otherwise, it is a clause. Phrases can be classified into six categories as follows.

Definition 23. A phrase in $\mathcal{L}$ is a combination of multiple words that forms an element of sentence in the categories of noun phrase (NP), verb phrase (VP), adjective phrase (AP), adverb phrase (DP), proposition phrase (PP), and complement phrase (CP) where each category of the phrases can be formally modeled as a specific set in $\mathcal{L}$, i.e.:

$$\begin{align*}
\text{NP} & \triangleq \{p_i \in \text{NP} \mid \text{NP} \subset |\mathcal{W} \subset \mathcal{L}| \} \\
\text{VP} & \triangleq \{p_i \in \text{VP} \mid \text{VP} \subset |\mathcal{W} \subset \mathcal{L}| \} \\
\text{AP} & \triangleq \{p_i \in \text{AP} \mid \text{AP} \subset |\mathcal{W} \subset \mathcal{L}| \} \\
\text{DP} & \triangleq \{p_i \in \text{DP} \mid \text{DP} \subset |\mathcal{W} \subset \mathcal{L}| \} \\
\text{PP} & \triangleq \{p_i \in \text{PP} \mid \text{PP} \subset |\mathcal{W} \subset \mathcal{L}| \} \\
\text{CP} & \triangleq \{p_i \in \text{CP} \mid \text{CP} \subset |\mathcal{W} \subset \mathcal{L}| \}
\end{align*}$$  \hspace{1cm} (21)

where $p_i$ represents a phrase, and $|\mathcal{W}|$ denotes a power set or combination of words in language $\mathcal{L}$.

Each category of phrases will be rigorously specified in the following subsections with formal models.
\textbf{Definition 24.} A noun phrase, NP, in language $\mathcal{L}$ is a noun composed with a determiner or adjective in front or following by alternative prepositional phrase and/or compliment phrase:

\[ NP \triangleq [r \mid r A \mid r AP] \mathcal{N} [PP] [CP \mid PP CP] \]

\[ (NP) \gamma NP \]

\[ (NP) \gamma ^{\prime} \]

where $\mathcal{N}$ is the principal noun including the subjective pronouns $X_{\mathcal{S}}$, objective pronouns $X_{\mathcal{O}}$, and gerunds $V_{\text{ing}}$ as defined in Definition 9.

\textbf{Definition 25.} A verb phrase, VP, in language $\mathcal{L}$ is a verb modified by an auxiliary, qualifier, adverb, adverb phrase, and/or negation in the following pattern:

\[ VP \triangleq [\alpha \mid \alpha \neg \mid \alpha \neg DP \mid \alpha DP \mid \alpha \neg DP] \mathcal{V} [DP] \]

\[ V_{\text{ing}} P = \text{to } V \]

\[ V_{\text{ing}} P = V_{\text{ing}} \]

\[ V_{\text{ing}} P = \text{BE } V_{\text{ing}} \]

\[ V_{\text{ing}} P = \text{HAVE } V_{\text{ing}} \]

\[ VP \gamma VP \]

\[ (VP) \gamma ^{\prime} \]

where $\mathcal{V}$ is the principal verb as defined in Definition 11.

\textbf{Definition 26.} An adjective phrase, AP, in language $\mathcal{L}$ is an adjective modified by an adverb, an adverb phrase, or a degree word in the following pattern:

\[ AP \triangleq [\delta \mid D \mid DP] \mathcal{A} \]

\[ (NP) \gamma AP \]

\[ (AP) \gamma ^{\prime} \]

where $\mathcal{A}$ is the principal adjective as defined in Definition 12.

\textbf{Definition 27.} An adverb phrase, DP, in language $\mathcal{L}$ is an adverb modified by a degree word $\delta$ in the following pattern:

\[ DP \triangleq [\delta \mid D \mid CP] \]

\[ DP (VP \mid DP \mid AP) \]

\[ VP DP \]

\[ (DP) \gamma DP \]

\[ (DP) \gamma ^{\prime} \]

where $\mathcal{D}$ is the principal adverb as defined in Definition 13.

\textbf{Definition 28.} A prepositional phrase, PP, in language $\mathcal{L}$ is a phrase of preposition with object that modifies a noun or noun phrase in the following pattern:

\[ PP \triangleq [DP] \mathcal{P} (\mathcal{N} \mid NP) \]

\[ (PP) \gamma NP \]

\[ (PP) \gamma ^{\prime} \]

where $\mathcal{P}$ is the principal preposition, and NP including the subjective pronouns $X_{\mathcal{S}}$, as defined in Definition 14.

\textbf{Definition 29.} A complement phrase, CP, in language $\mathcal{L}$ is an adjective or noun phrase that completes the sense of a verb, a subject, and an object in the following pattern:

\[ CP \triangleq A[\mathcal{N}] \mid AP [NP] \]

\[ (CP) \gamma AP \]

CPs are very widely used phrases in sentences, which can be a supplemental part of N/NP, V/VP, A/AP, or P/PP [Hodges and Whitten, 1982; O’Grady and Archibald, 2000; Taylor, 2002].

\section*{4. The Formal Sentence Structures of English}

On the basis of the syntactic structures, mathematic models, and grammar rules described in Section 3, a formal model of sentence in English can be rigorously developed in this section.

\subsection*{4.1 The Formal Model of Sentences}

A sentence is a basic unit of a complete expression for a thread of thought or a behavior. The general structure of a sentence in a natural language encompasses a subject and a predicate where the latter can be further divided as a verb phrase (behavior) and its object(s).

\textbf{Definition 30.} An abstract sentence, $S$, in language $\mathcal{L}$ can be formally described as a triple:

\[ S \triangleq [\mathcal{J}] \mathcal{V} [\mathcal{O}] \]

where

- $\mathcal{J}$ is the subject of the sentence, $\mathcal{J} \in (N \cup X \cup \mathcal{N} \cup \mathcal{O}) \subset \mathcal{N} \subset \mathcal{W} \subset \mathcal{L}$;
- $\mathcal{V}$ is the predicate of the sentence with an inflective form of verb or VP, $\mathcal{V} \in \mathcal{V} \subset \mathcal{W} \subset \mathcal{L}$;
- $\mathcal{O}$ is the object of the predicate and the sentence, $\mathcal{O} \in (N \cup X \cup \mathcal{N} \cup \mathcal{O} \cup \mathcal{X}) \subset \mathcal{N} \subset \mathcal{W} \subset \mathcal{L}$, where the brackets [ ] represent an optional term.

Sentences may be classified according to their purposes known as statements, commands, questions, and exclamations. Sentences may also be classified according to their syntactic structures known as simple and complex ones. A complex sentence is usually constructed by conjunctions of subsentences known as clauses.

\textbf{Definition 31.} A clause is the third-level structures beyond a phrase that is a subsentence related to another one in a complex sentence usually connected by a coordinative or
correlative conjunctions as in independent clauses and introduced by subordinate or relative pronouns as in dependent clauses in the following pattern:

\[ S \triangleq (\text{Independent clause (IC)} | \text{Dependent clause (DC)}): \]

\[ = IC = (C, C) \]

\[ | (C, C) \]

\[ | (C \gamma_{\text{Coordinating}} C) \]

\[ | (C \gamma_{\text{Correlative}} C) \]

\[ ) \]

\[ DC = (C_m \gamma_{\text{Subordinate}} C_s) \]

\[ | (C_m X_f C_s) \]

\[ ) \]

where \( X \) represents relative pronouns such as the general personal (\( X_p \)), indefinite (\( X_i \)), demonstrative (\( X_d \)), and interrogative (\( X_q \)) pronouns, respectively, as defined in Eq. 8.

The connections between two clauses in a complex one can be established by the conjunctions (\( \gamma \)) as described in Definition 21. The relations of clauses in a complex sentence can be equivalent and subordinate. In the former case, the clauses are known as independent clauses (IC); while in the latter case, the clauses are known as the dependent clauses (DCs) where the main clause is in the form of a complete sentence, and the subordinate clause is a sentence introduced by a conjunction or relative pronoun.

There are noun, adjective, and adverb clauses from the syntactical point view. In some cases the introductory conjunction for a subordinate clause may be omitted when the syntax and semantics are reserved and be clearly understood. This form of clause is called elliptical clauses.

4.2 The Deductive Grammar of English

On the basis of the mathematical models of both the individual syntactic structures as developed in Section 3, a formal model of the English grammar known as the deductive grammar can be rigorously established as shown in Fig. 1.

**Definition 32.** The deductive grammar is an abstract grammar that formally denotes the syntactic rules of a language based on which as a generic formula, valid language sentences can be deductively derived.

The Deductive Grammar of English (DGE) can be formally described in EBNF as shown in Fig. 1. DEG specifies the formal structures of sentence, clauses, phrases, modifiers, and terminals (lexes), as well as their syntactic rules, in a top-down hierarchy. Based on the formal DGE and supported by a lexical database for all terminals words, a parser can be derived to autonomously process texts expressed in English according to the formal grammar.

The syntactic rules of individual elements such as those of lexes, modifiers, and phrases as well as their inflections have been formally defined in Section 3. In addition to the syntactic rules, there are 15 relational semantic rules between different elements of sentences in English such as those of \( \beta \)-agreement, \( \gamma \)-agreement, negation, and inverted structures [Wang & Berwick, 2012].

![Fig.1 The deductive grammar of a generic sentence in English](image-url)
by Table 2 where the higher level structures describe the syntactical models of a general sentence in DGE; while the bottom level structure is the general pattern of sentence, which function as a template for deductively deriving any concrete instance according to DGE. The DGE can also be illustrated by a syntactic diagram for the general sentence pattern as the most general and most complicated structure of English syntax.

A number of examples are demonstrated in Table 2 against the general pattern of sentences in English. Observing Fig. 1 and Table 2, it is noteworthy that the syntactic structure of the DGE pattern is highly recursive, particularly all forms of phrases in the subject, object, and predicate of the sentence. The recursive property of elements in sentences at the levels of parts of speeches, phrases, and clauses is interesting demonstrated in the general pattern, particularly the recursive structure of clauses as shown in Example 3 by clause $\ell_3$. In order to save space, a few of simple phrases in Table 2 are not completely extended to its terminal lexical elements. The case studies indicate that no matter how simple or complex an instance sentences would be, they can be efficiently and effectively recognized, analyzed, and processed by the general formal model of DGE by human or machines.

<table>
<thead>
<tr>
<th>Syntactical Structure</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence</strong> ($S$)</td>
<td>$\tau$</td>
<td>$\tau$</td>
<td>$\tau$</td>
<td>$\tau$</td>
</tr>
<tr>
<td><strong>Subject</strong> ($\ell$)</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td><strong>Predicate</strong> ($\ell'$)</td>
<td>$\beta$</td>
<td>$\beta$</td>
<td>$\beta$</td>
<td>$\beta$</td>
</tr>
<tr>
<td><strong>Object</strong> ($\ell''$)</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td><strong>NP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>VP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>End</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Table 2. Application Examples of the Deductive Grammar of English

<table>
<thead>
<tr>
<th>Syntactical Structure</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence</strong> ($S$)</td>
<td>$\tau$</td>
<td>$\tau$</td>
<td>$\tau$</td>
<td>$\tau$</td>
</tr>
<tr>
<td><strong>Subject</strong> ($\ell$)</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td><strong>Predicate</strong> ($\ell'$)</td>
<td>$\beta$</td>
<td>$\beta$</td>
<td>$\beta$</td>
<td>$\beta$</td>
</tr>
<tr>
<td><strong>Object</strong> ($\ell''$)</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td><strong>NP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>VP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>End</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. Conclusions

This paper has represented a formal, general, and precise methodology for natural language processing in cognitive and computational linguistics. Denotational mathematics has been adopted to model and specify the Deductive Grammar of English (DGE). The work has demonstrated that the grammar of a natural language can be formally specified by a set of rigorous syntactical structures and rules. Based on them, cognitive systems for machine-enabled language processing and learning may be implemented. A wide range of applications of the formal models of DGE have been identified in natural language processing, computational linguistics, and cognitive informatics in general, and in word processing, web search engines, machine language comprehension, autonomous machine learning, smart cell phones, semantic computing, cognitive computing, and computing with words in particular.

Acknowledgement

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


