Generate of Construction Tree based on HPSG for Euclid "Elements"

Keisuke Senda (010062)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 2002

Keywords: HPSG, syntactic analysis, LiLFeS, ancient Greek.

In this paper, we describe Head-driven Phrase Structure Grammar (HPSG) parser for ancient Greek we developed. The purpose of this work is to help to work about classics.

In research about classics, we can use books, which are manuscripts of target literature. It is rare that we can use original. Are these books genuine or fake ? The text appearance is used about the judgement. However, this judgement is difficult. We need to judge not only by the text but by also other texts of the some author, as well as those of other authors. Therefore, the number of texts are enormous and there are many differences between the original and the manuscripts. In manuscripts, transcribers might have make errors, so text appearances of the original become ambiguous. In addition, revisions may change the original text. Therefore, important information is buried in not important information.

What is more, in computerized preservation of classics, it is prevalent to preserve only the text. These texts play an important role because of easiness of searching words. However, to add other information makes the text more available. We think syntactic structure is one of such information, which will help to find differences among manuscripts.

"Elements" written by Euclid, is one of such classical literature. He was a mathematician in ancient Greece, and described this literature around B.C. 300. "Elements" might have the problem we mentioned above. It

is references in propositions. In "Elements", when refers to propositions appeared before, it don't refer to the number of proposition. Instead, it repeats the proposition's expressions. Therefore, it is hard to search the referred parts. We considered that it is possible to solve this problem to use the information of construction tree. Thus, we decide to parse "Elements".

Phrase structure grammars, as regular grammar and context free grammar, are one of typical grammars for parsing. However, the phrase structure grammar is difficult to get semantic information. Therefore, we adopt a unification grammar. Unification grammar can analyse syntactic and semantic structure at once. Our purpose is to add semantic information so as to treat classics in handy. Among various unification grammars, we adopt HPSG. HPSG has rich information in lexical entries, and the number of grammatical rules (ID-schemata and principles) are small. This character is adequate for Greek, because the language has many conjugations. For example, noun has gender, number and case, and verb has voice, mood, person, number and tense. If we adopt CFG, we must write so many rewriting rules as conjugations. Those rules can be adjusted in HPSG by adding lexical entries. It is easier than to add rewriting rules, because to add rewriting rules may cause errors at other sentences. The number of words which appear in "Elements" is not so many, about 300 to 500 words. In these points, we conclude that we adopt HPSG is better-suited.

Consequently, in this research, we implemented the system that construct parse trees for "Elements". And, we estimated the feasibility of this system. This system was implemented with "LiLFeS", a programming system for linguistic formalisms based on typed feature structures. HPSG's lexical entries and grammatical rules are expressed by typed feature structures. The core engine of LiLFeS is an abstract machine developed for efficient handling of typed feature structures. On LiLFeS, there are HPSG parsers for English and Japanese grammar with wide coverage. Therefore, we construct grammar for Greek, and parse "Elements".

The first, we expanded feature structure for Greek. This expansion is mainly used to correspond to Greek word's conjugations. Herewith, all words in Euclid "Elements" are written by typed feature structures. Next, we improved ID-schemata based on it. In Greek, orders of words (phrases) are comparatively free, because words have information of case. So, en-

tering information of case in feature structure and disregarding order of words make parsing possible.

But, these works are incomplete to parse "Elements". Because, Greek has words referred to as "particle". Particles work like conjunction or adverb. They appear at the second position of sentences. It is difficult to solve this problem with parsing. So, we committed the solution to preprocess. As a result, the preprocess shifts the position of particles, and we can parse the sentences. Other works of preprocessor are mentioned below : adjustment of spelling, adjustment of comma, adjustment of order of articles and the other words, and supplement of omitted words. In these preprocesses, complement of omitted words are difficult to extrapolate. Thus, partial supplements are manual labor.

As we mentioned before, we implemented the preprocessor and the grammar for ancient Greek. The preprocessor helped parsing at various adjustments and supplements. In parsing, this grammar was expanded by Greek's grammar. Therefore, we could generate lexical entries with detail information, and ID-schemata were expanded. By the way, the number of lexical entries is 229, and the number of ID-schemata is 10. As a result this grammar can generate parse trees for 75 sentences of 76 sentences in "Elements" Book 2 Demonstration 2 to 5.