An English Grammar Checker as a Writing Aid for Students of English as a Second Language

Jong C. Park[1], Martha Palmer[1], and Gay Washburn[2]

[1] Computer & Information Science
University of Pennsylvania, PA 19104
{park,mpalmer}@linc.cis.upenn.edu

[2] English Language Programs
University of Pennsylvania, PA 19104
gwashbur@sas.upenn.edu

1 Introduction

We present a prototype grammar checker for English as a Second Language (ESL) students, utilizing Combinatory Categorial Grammar (CCG) written in SICStus Prolog. Instead of attempting to handle all possible grammatical errors, the grammar checker identifies certain specific types of grammatical mistakes that appear more regularly than others in the present domain of application.

2 Grammatical Mistakes

The current project started as part of a collaboration between the Computer and Information Science department and the English Language Programs at the University of Pennsylvania in an effort to provide a computational tool for students who are learning English as a second language. As an initial attempt, the present implementation focuses on certain dominant types of syntactic mistakes, as identified from the available set of students' essays.

The kind of mistakes that are detected by the current system includes: Wrong capitalization (sentence initial, wrong lowercase/uppercase initial letter), missing fragments (subjects, objects, some prepositions, complements, articles, clauses, *the*, *than*, etc), some extra elements (e.g., the infinitive marker after auxiliary verbs), wrong agreement (number, case, etc), wrong verb form, and various mismatches (verb tense with adverbs, etc).

3 Grammar Formalism

In order to handle ungrammatical expressions as well as grammatical ones, we utilized the syntactic domain of locality implicit in categorial lexical entries under the CCG framework, by including known variations of categorial association in the lexical entries. Multiple parts-of-speech are handled by putting all the assertions regarding grammatical parts-of-speech before those regarding ungrammatical ones.

4 Categorial Lexicon

In constructing a categorial lexicon for the grammar, we used an existing grammar-independent lexicon that had been made available as part of the XTAG project at the University of Pennsylvania. The original public domain lexicon contains about 37K quintuples, (INDEX, ENTRY, POS, CAT, FS), where POS and CAT are associated with a part-of-speech (such as V or N) and a set of categories, respectively, for the lexical item associated with ENTRY. The lexical items are supplemented by a morphology table, which has about 317K entries for various grammatical inflections. We have also made use of public domain word lists and consulted an on-line electronic dictionary for more commonly used lexical items.

At the moment, the lexicon has about 18K nouns, 6.5K adjectives, 4.5K transitive verbs, 2K intransitive verbs, among other categories.

5 English Grammar Checker

The user interface of the system includes a direct interaction with the Prolog interpreter, as well as an Internet interface. The Internet interface is written in HTML and CGI/PERL, which invokes the runtime image of the Prolog code.

The project is still at an initial stage and there are many other issues that need to be addressed, such as performance and coverage of the mistake types. We believe that the use of a standard linguistic theory such as CCG made it possible to develop a grammar checker in a very short time frame and limited man power, as the existing large standard lexicons are be made readily available for it. Pending further funding, we believe that the prototype grammar checker can be brought up to an industry strength.

Acknowledgements