TRANEL (Travaux neuchâtelois de linguistique), 1994, 21, 223-228.

The ARCTA Prototype: An English writing tool and grammar checker for French-speakers ◊

Corinne Tschumi, Franck Bodmer, Etienne Cornu, François Grosjean, Lysiane Grosjean, Natalie Kübler, Cornelia Tschichold[1]

Abstract

This paper presents a second language grammar checker for French native speakers who write in English. It was developed at the Language and Speech Processing Laboratory of the University of Neuchâtel (Switzerland) and includes on-line writing tools (dictionaries, grammar helps, a translation tool for set expressions, a verb conjugator, etc.) and a grammar checker that corrects morphological, lexical and syntactic errors. False friends and other potential errors are dealt with in such a way as to reduce overflagging. Particular attention has been paid to second-language errors (as opposed to the ones made by native speakers) and to user-friendliness.

1. Introduction

The research presented here corresponds to a three-year project funded by the CERS (Swiss Committee for the Encouragement of Scientific Research) in which the Language and Speech Processing Laboratory of the University of Neuchâtel developed a new second language grammar checker for French native speakers who write in English. This article gives a general overview of the whole project. More specific aspects are developed in the next three articles. In what follows, we first present the basic philosophy behind our prototype (called ARCTA, which stands for "Aide à la rédaction et à la correction de textes anglais"). Next we explain how we analyzed a corpus of English texts written by French-speakers in order to identify and classify typical second-language (L2) errors[2]. Then, we focus on the different kinds of on-line tools included in the prototype and we present a special

◊ This research was supported by a grant from the Swiss CERS/KWF (2054.2).
1 We would like to thank Jacqueline Gremaud-Brandhorst, Nicolas Lévy, Catherine Liechti, Tracy Mannon, Alain Matthey and Ann Morel who, in one way or another, contributed to this project.
2 See the article by C. Tschumi & C. Tschichold (this issue).

mechanism used for treating potential errors such as false friends. Finally, we describe how grammar checking takes place, from text segmentation through word disambiguation[3] and island syntactic pre-processing to error detection[4] and correction.

2. The philosophy behind the prototype

The aim of the ARCTA prototype is to facilitate the work of French-speakers when writing English. In developing the prototype, we have tried to keep in mind the following basic principles (although we have sometimes found we could not always apply them for technical reasons):

a) detect and correct as many errors as possible;

b) limit overflagging (false alarms) to a strict minimum;

c) concentrate on the most frequent mistakes;

d) only deal with widely accepted mistakes;

e) interact with the user when necessary (but not too often);

f) have a user-friendly interface;

g) make use of up-to-date technology;

h) use linguistic databases that can easily be updated.

Limiting overflagging (point b) is particularly important in the context of L2 texts as stopping on a non-error may mislead the user whose mother-tongue is not English. In addition to helping users write a text and proof it, we have given them the opportunity to learn more about the use of English and improve their command of the language. To do this, we have offered them various help options during the writing process and the checking phase.

[3] See the article by F. Bodmer (this issue).
[4] See the article by N. Kübler & E. Cornu (this issue).

3. Corpus analysis and error typology

At the onset of the project, we made up a typology of errors produced by French-speakers in English. To do this, we used our own knowledge of students' mistakes, printed lists of errors in books or articles and the errors predicted by English-French comparative linguistics (Guillemin-Flescher, 1981; Kübler, 1992; Vinay & Darbelnet, 1977). In addition, and more importantly, we analyzed a corpus of texts (some 27,000 words) taken mainly from high school and business school written exams. The corpus was carefully analyzed and corrected by three native speakers of English (who all teach ESL) and the 2,862 errors found were used to complete our typology as well as give us the frequency of each error type. The corpus not only gave us a large number of errors but it also allowed us to work on real text with all of its inherent difficulties.[5]

4. On-line writing tools

A poll conducted among potential users helped us identify the various on-line tools that they would welcome. We finally chose the following for our prototype:

a) an English monolingual dictionary,

b) two bilingual dictionaries (French-English and English-French),

c) a verb conjugator,

d) grammar helps,

e) information on difficult words,

g) a translation tool for set expressions.

The above tools have not yet been developed exhaustively but the access mechanisms together with a minimum amount of linguistic data have been included in the prototype.

[5] For a more detailed presentation of this aspect of our work, see the article by C. Tschumi & C. Tschichold (this issue).

5. Potential errors

After analyzing the errors in our typology and taking into account the current state of computational linguistics, we acknowledged, like others have done before us (Payette & Hirst, 1992; Thurmair, 1990), that some mistakes could not be detected automatically. These include false friends, confusions and other lexical difficulties where an understanding of the meaning of the sentence (or paragraph) is necessary to determine whether there is a mistake or not. The problem with most traditional grammar checkers is that they flag all occurrences of these potentially problematic words without being able to determine whether they are indeed used wrongly, thus triggering many annoying false alarms. In order to avoid this overflagging but to include these categories nevertheless, we decided to treat them separately. Thus, the user has a special "potential errors" check which lists all the potentially problematic words in the text. When he clicks on one of the items in the list, it is highlighted in the text and an information window appears which warns the user about the specific difficulties attached to that word.

6. Error detection and correction

Error detection and correction is done in several steps. First the text is segmented into sentences and lexical units. Spelling mistakes are corrected with a commercial spell-checker and then each word is looked up in a dictionary and receives a tag for each of its possible syntactic categories. (We are currently using an extract of CELEX that contains all the words in our corpus plus the most frequent words of English, totalling about 4,800 canonical entries.) The tagged text then goes through a word class disambiguator which gives each lexical unit only one syntactic category. This disambiguator is based on neural networks.[6]

In the next stage, syntactic analysis is undertaken. Since we are dealing with L2 texts that contain a large number of errors at all linguistic levels, a complete syntactic parse of the sentence is not an option. Instead, we have chosen island processing, starting with the identification of simple noun phrases with an NP parser based on stochastic methods.

[6] See the article by F. Bodmer (this issue).

The main phase of processing is achieved with finite state automata (Winograd, 1983). At the pre-processing level, automata group simple NPs into complex NPs, mark temporal NPs, locate the head of an NP, identify prepositional phrases such as temporal PPs and analyze verb forms for tense, mode, aspect and voice. This information is used by the next level of automata - the detection automata - which are specifically geared towards detecting erroneous sequences of words, morphological errors, problems of agreement (subject - verb, or within an NP), etc. Sometimes a single automaton cannot detect an error and we have to resort to filter automata before the actual detection automata. For instance, filter automata are used when we want to detect an error in different contexts but want to give the same correction message, or when we need to locate a correct sequence in order to eliminate it before detecting an error.[7]

Whenever there is insufficient syntactic or semantic information, the user is asked for some information, such as "Are the words X-Y the subject of the verb Z?" Depending on the answer, we proceed with the detection or move on to another problem. Despite the fact that the user interaction option is both attractive and useful, we have tried to use it sparingly as it might bother the user a bit too much.

The user is given a solution for correcting an error whenever possible but this is sometimes difficult, as we have not developed a full morphological analyzer. When error correction can take place automatically, we offer to either replace a word or several words with others, add a specific word (such as a missing article or preposition), delete one or several words, or permute two (groups of) words. When it is not possible, a message is produced stating that a particular mistake has been found and explaining, usually with an example, how the text can be corrected manually. An on-line grammar is available during the correction process. It is important to note that users remain in charge of their text at all times. They have the choice of accepting a proposed correction or not and of editing their text at any given moment.

[7] See the article by N. Kübler & E. Cornu (this issue).

We have put a lot of emphasis on giving simple yet accurate messages. This is particularly important during the interaction phase where the question needs to be clear and precise and yet not contain grammatical jargon that is incomprehensible to a non-linguist.

The prototype runs under Windows and the interface has been developed with Visual Basic.

7. Conclusion

Our prototype is the result of a careful analysis of the errors that are made by French-speakers when they write in English, of appropriate computational techniques and of a clear assessment of what a user needs and wants while writing and subsequently proofing his or her prose. Hopefully, it will enter a commercial development stage in the near future.

8. Bibliography

GUILLEMIN-FLESCHER, J. (1981): *Syntaxe comparée du français et de l'anglais*, Paris, Ophrys.

KUBLER, N. (1992): "Verbes de transfert en français et en anglais", *Linguisticae investigationes*, 16 (1), 61-97.

PAYETTE, J. & G. HIRST (1992): "An Intelligent computer-assistant for stylistic instruction", *Computers and the Humanities*, 26, 87-120.

THURMAIR, G. (1990): "Parsing for grammar and style checking", *COLING*, 365-370.

VINAY, J.P. & J. DARBELNET (1977): *Stylistique comparée du français et de l'anglais*, Paris, Didier.

WINOGRAD, T. (1983): *Language as a Cognitive Process: Syntax*, Reading, Mass., Addison-Wesley.