Publié dans Revue Tranel (Travaux neuchâtelois de linguistique) 21, 229-234, 1994
qui doit être utilisée pour toute référence à ce travail

Selecting English errors made by French-speakers for automatic correction ◊

Corinne Tschumi & Cornelia Tschichold

Abstract

In this article, we first present how we went about defining what we were going to treat as an error for detection and correction by a grammar checker of English for French-speakers. We then explain how we found errors--in great part through an in-depth analysis of a corpus of English texts written by French-speakers--and how we classified them in order to come up with a typology. Finally, we list the criteria that were used to select the errors that would be considered for automatic detection and correction.

1. Introduction

The work we present here is part of the ARCTA Prototype project which consists in the development of a second language grammar checker for French native speakers writing in English[1]. In this research, we deal with second language errors (L2 errors), more precisely errors in English texts written by French-speakers. These L2 errors are different from L1 errors in English (made by English native speakers) in that certain typical monolingual mistakes do not appear whereas other types of mistakes do. Since English is not the users' mother tongue, certain mistakes, which are due to pronunciation for instance, are rarely made (e.g. the confusion between "you're" and "your"). On the other hand, there are mistakes which a native speaker would not usually make, for example:

- lexical confusions: **Chess is a slow play*
(the French noun *jeu* has two meanings: *play* or *game*),

- difficult syntactic constructions: **The most he eats, the hungriest he is*

◊ This research was supported by a grant from the Swiss CERS/KWF (2054.2).

[1] For a general presentation of the project, see the article by C. Tschumi, F. Bodmer, E. Cornu, F. Grosjean, L. Grosjean, N. Kübler & C. Tschichold (this issue).

- problems of word order: **in the few last years*, etc.

At this stage of selecting errors, our work consisted of four different tasks which are developed in the sections below:

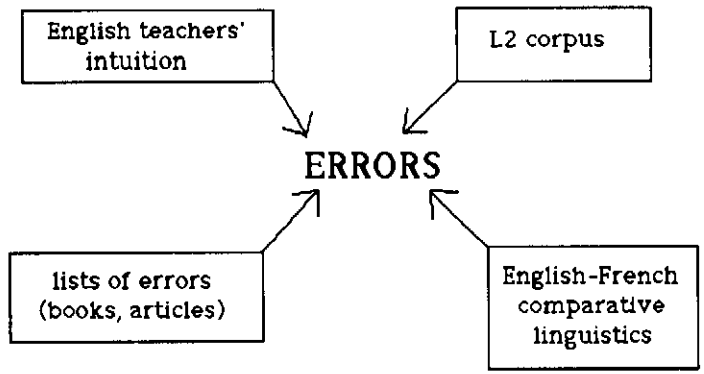
- defining what we were going to consider an error,
- finding the typical errors made by French-speakers in English,
- classifying these errors,
- selecting the errors that can be detected (and, if possible, corrected) automatically.

2. What is an error?

A linguistic error in a language is what is wrong according to a grammatical norm (e.g. **Why did she wrote this letter?*) or according to a given register (*It ain't too late* is not acceptable in a business letter, whereas it can very well be part of a dialogue in a novel). An error will usually strike a native speaker as being anomalous. This naturally led to the question of the norm we wanted to consider, since our definition of an error depended on it. First we had to choose between written and oral language. Then we had to opt for a geographical norm: British, American, or yet another. We finally chose the written American norm. The choice of a written rather than an oral language norm was easy in the context of written texts only. Since we were dealing with written L2 texts, we adopted a standard norm, excluding colloquialisms that are too informal or at times vulgar.

3. Collecting errors

Having decided on the norm we wanted to use, the question became: how do we go about finding typical errors made by French-speakers when they write in English? We identified four possible sources, as can be seen in the diagram below:



- using the intuition of teachers of English who are familiar with their students' typical mistakes,
- consulting published lists of French-speakers' mistakes,
- collecting a corpus of texts written by French-speakers and correcting them,
- trying to predict errors on the basis of English-French comparative linguistics.

There are, however, problems with these approaches. The problem of **completeness** is a question that arises with all four approaches. Appealing to the intuition of English teachers raises the question of **subjectivity**. Is the error actually an error? Could it be that, on the other hand, a sentence accepted by the teacher contains an error? The analysis of a corpus leads to other questions: **What kinds of texts** should be examined and **how many**? **What competence level in English** should be considered? Examining lists of common mistakes also raises many questions, since the precise source of the errors is often not specified. Errors mentioned in these lists might have been taken from a corpus or they might have been constructed by the authors using their own intuition or have come from yet another source. Finally, comparative linguistics itself is not an ideal solution either, because a

grammatical point that is treated differently in French and in English does not necessarily lead to an error (especially if well-drilled at school) [2].

The approach we adopted to collect errors consisted in combining the various methods described above. We started by mainly using English teachers' intuition of French-speakers' mistakes to which we added published lists of errors. This enabled us to compile a first draft of the typology. (The actual classification of the errors is dealt with in the next section.)

We then collected a corpus of texts (containing some 27,000 words) written by native speakers of French, more precisely by students who had studied English for about 4 to 6 years. In order to have a variety of texts, we used *maturité* exams from both the *gymnase* and *école de commerce* (high school level) as well as some university exams (*demi-licences*). There were four different types of texts, namely essays, answers to questions, summaries and translations.

Three ESL teachers, all English native speakers, were given the texts and asked to correct them separately. The results were used to revise and complete our typology and gave us information regarding error frequency.

4. Classifying errors

We built our typology of errors around seven main categories:

1) Errors related to the **graphical form of words**. In this category we included, among others:

a) spelling mistakes:

- interferences with French (**adress*, *marchant*),
- non-words (**buisness*, **the ferst time*), [3]

[2] See for instance H. Dulay & M. Burt, "Errors and strategies in child second language acquisition", *TESOL Quarterly*, 8(2), 1974.
[3] Since our corpus was originally written by hand, we do not have typing mistakes. These are usually quite frequent in typed texts.

b) morphological errors (**certains childrens*),

c) mixtures of British and American spellings (? *labour and honor*).

2-5) Errors related to different syntactic categories: **adjectives, adverbs, nouns, verbs.**

Within each of these four categories, we subdivided errors according to morphology, lexicon, syntactic aspects (words that often precede or follow the category in question), punctuation, agreement, tense, and so on.

6) **Combinations of words** which include phrases, collocations, dates, idioms... (**He seized the occasion*, **on the 17th June*, **Arrange them by alphabetical order*, **He never does his bed*)

7) **Utterance** errors including negations (**She came not*), reference pronouns (**If the firm calls, tell her...*), word order (**He asked me when are we leaving*), agreement (**There is people who think...*), subordinate clauses (**He was sent abroad for learning Spanish*, **This is all what is left*), prepositions (**I have been reading since two hours*), syntactic constructions (**They will come, don't they?*), etc.

5. Selecting errors for detection and correction

The selection of a specific error does not necessarily mean that the error will actually be **corrected** by the grammar checker. There are in fact two phases, or levels of processing, of an error: first, detecting the error (sometimes with the help of the user), and secondly, finding and proposing one or several corrections (again perhaps with the user). Some errors will only be covered in the first stage--leading to a warning message to the user--while others will be defined well enough to result in one or more propositions for correction.

Whatever the final treatment of the error (detection only, or detection and correction), several criteria were taken into account when selecting the errors that were to be covered. One important factor was the **frequency of the error** which we derived from the corpus analysis. An equally important factor concerned the **computational complexity** involved in error detection and correction. This of course depends on the

computational tools chosen (in our case, a system based on automata for both error detection and a local grammatical analysis[4]). Computational complexity, which comes down to computational feasibility, is bordered by not detecting an error ("miss") on the one hand and flagging a non-error ("false alarm") on the other. Other factors that we took into account were the **impact on comprehension** of the error (Does the error make the sentence almost impossible to understand or can we interpret the message despite it?) as well as the **scientific interest of the error** (How is this error covered by other grammar checkers? If they can deal with it appropriately, is it worth spending much time on it, especially if we cannot do any better?). Finally, the last factor concerned the **needs of potential users** expressed in the responses to a survey we conducted.

These factors provided us with a good basis for carefully selecting the errors we would attempt to detect and correct.

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