Analyzing a corpus of documents written in English by native speakers of French:
Classifying and annotating lexical and grammatical errors

Camille Albert
Cultures Anglo-Saxonnes
Université Toulouse Le Mirail
calbert133@aol.com

Marie Garnier
Cultures Anglo-Saxonnes
Université Toulouse Le Mirail
mhl.garnier@gmail.com

Arnaud Rykner
Lettres, Langages et Arts
Université Toulouse Le Mirail
arnaud.rykner@neuf.fr

Patrick Saint-Dizier
Institut de Recherche en Informatique de Toulouse
CNRS
Université Paul Sabatier
stdizier@irit.fr

Abstract

In this paper, we present a work in progress whose final objective is to design an automatic error correction tool using a framework of argumentation in order to explain errors to the user (project *CorrecTools*). This work is conducted in a strong didactic perspective. We focus on the analysis and correction of errors produced by French speakers writing in English, as errors are often particular to one community of speakers. As a first step, we construct a corpus of heterogeneous and representative productions ranging from emails to scientific publications. Lexical and grammatical errors found in this corpus are classified according to a system based on linguistic criteria. Errors are then annotated using XML annotations, which we also use to make correction proposals, and eventually draft correction rules.

1. Introduction

Documents produced in a language other than their authors' native language often contain a number of typical errors that can make comprehension difficult. The aim of our project, dubbed *CorrecTools*, is to categorize such errors using pairs of languages (in this paper, French authors writing in English), and to propose a number of correction strategies which can be implemented in such devices as text editors and email front-ends. We focus on those types of errors which are not treated by advanced text editors. Considering pairs of languages greatly facilitates the proposition of corrections, due to errors being generally prototypical.

Using a framework of argumentation, we intend to develop a system which is able to explain errors to the user, issue advice as to the proper correction, and provide additional grammatical and lexical information. Through user profiling, the system should adjust to the level and the needs of the individual user.

One of the objectives of this research project is to explore the cognitive strategies used by human correctors (e.g. teachers) as they detect and correct errors. These strategies are used

to design our annotation system, which aims at describing errors as segments with their own specificities rather than simply as grammatical deviancies.

In this paper, we present the first step of this research project. First of all, we introduce our method for the constitution of an exploratory corpus, i.e. the different parameters taken into consideration and how they are realized in the corpus. Then we put forward the classification system that we have designed in order to categorize errors, as well as a synthesis of the difficulties encountered and the improvements considered. Finally we present a system for the annotation of errors and the proposition of corrections.

2. Constructing a corpus

The constitution of a corpus is a fundamental step when attempting to analyse errors. Indeed, the parameters taken into consideration, and thus the nature of the corpus, determine to a large extent the type of errors that are going to be discovered, as well as the system of categories that is going to be created. One of the main difficulties is to determine the types of documents that enter into the composition of the corpus, since it should contain the errors which are encountered most frequently in the productions of French speakers writing in English.

2.1 Methodology

At the stage of preliminary analysis, we chose to focus on publications and emails, which were relatively accessible documents and were representative of situations implying the use of English in order to communicate. We observed a great disparity in the types of errors found in emails and publications. So as to widen the scope of the initial corpus, we included other types of documents in the form of official and personal web pages and blogs. These documents combine the characteristics of several types of documents, since they are designed for perusing on the internet while very often adopting a formal style. We were thus able to observe a wide range of errors produced by French speakers writing in English in natural situations.

This preliminary study enabled us to identify a number of fundamental parameters to take into account in the construction of our exploratory corpus, as these parameters would determine the type of errors observed. These are listed below:

- register (familiar or formal expression)
- level of control (amount of care devoted to the production of a document)
- field of document production (business, research, personal sphere, etc.)
- type of authors (professionals, researchers, students, etc.)
- target audience (business partners, hierarchy clients, program committees, friends and family, etc.).

Some of these parameters seem to be consistent with certain types of documents, and therefore they are often found together. For example, a document written with a low level of control might also contain familiar expressions, and be targeted at family members or friends. On the contrary, a document targeted at a program committee is usually written with formal expression and a high level of control. However, this redundancy is not systematic: professional emails often manifest a low level of control while having been written in a formal style.

As a complement to the type of documents already included in our corpus, we intend to investigate the possibility and relevance of using learner corpora. Since the tool that we want to design is not particularly targeted at students of English nor especially designed to be used as a teaching aid, the use of this type of corpora is not a straightforward choice. However, it might

be a satisfactory solution to some of the problems encountered in the constitution of our corpus, such as the scarceness of errors in articles and the overwhelming number of errors in emails, which makes their manual detection, annotation and correction very time-consuming. We have the opportunity of using already-existing learner corpora, such as the *International Corpus of Learner English* (Granger et al, 2009), as well as to compile our own learner corpora.

2.2 Parameters

The following parameters are the main ones that were taken into consideration in the construction of our exploratory corpus:

- diversity of authors: we have gathered documents from 60 authors (35 for productions with a low level of control, 25 for productions with a high level of control); we estimate the numbers of authors in the case of websites to be about 19 (it is difficult to give a precise number of authors in the case of such documents), and the level of control varies according to the source of the website;
- diversity of fields or domains: the documents of our corpus come from a number of fields, such as public services, business, tourism, scientific research, etc. This diversity allows us to take into account different linguistic habits and modes of expression which are proper to some communities
- diversity of documents and levels of control: in order to be representative of the most frequent textual errors, our corpus includes about 140 pages of text (90 pages for low control level productions; 50 pages for high control level productions). As explained in Section 2.1, we attempted to take into consideration a wide panel of easily accessible documents: emails, blogs, forum posts, scientific publications, reports etc. Most documents that come from the internet (emails, blogs, forum posts and personal web pages) are associated with a low level of control, whereas publications, reports and professional web pages correspond to a high control level. However, this parameter varies a lot from one source to the next, and we observe the existence of a continuum between the two poles representing high control level and low control level productions.

The characteristics of the exploratory corpus thus constructed, and defined for a feasibility study, are summarized in the following tables:

Level of control :

Level of control / Type of document		High	Average	Low
Publications		×		
Websites	Personal			×
	Tourist Information Office		×	
	Gastronomy		×	
	University	×		
	Hotels		×	
	Public administration	×		
Emails	University		×	
	Computer sciences			×
	Aeronautics			×
	Medecine			×

Table 1.a

Authors :

Document type	Size (number of words)	Number of authors
Publications	33564	25

Websites	7694	19 (estimation)
Emails	9331	35

Table 1.b

Fields :

Document type	Fields	Size (number of words)
Publications	University	33564
Websites	Personal	457
	Tourist Information Office	2194
	Gastronomy	1836
	University	2380
	Hotel	601
	Public administration	226
Emails	University	2097
	Computer sciences	1753
	Aeronautics	3018
	Medecine	2282
	Family	181

Table 1.c

3. Classifying errors

We create categories of errors according to their nature rather than according to the type of corrections that they should receive. In this preliminary study, we have chosen to base our classification on linguistic criteria rather than on the observation of surface phenomena (i.e. omission, addition, misuse, wrong order etc.) (Ellis, 1994). Erroneous segments are grouped according to the syntactic phrase they constitute or are a part of (i.e. noun phrase, prepositional phrase, verb phrase, sentence or clause, etc.). This ensures that this classification is understood by as many end users and annotators as possible, while enabling structural as well as semantic errors to be taken into consideration at the level of the phrase or the clause. Moreover, this system yields categories of the same linguistic rank, which guarantees a certain degree of internal coherence. It can also be used to study other pairs of languages.

The finer distinctions that are included inside main categories were obtained through the observation of the type of errors found in the corpora. As the latter expands, this second level of categories is bound to evolve accordingly. At the moment, we also acknowledge a main category concerning “other” errors (mainly lexical errors), which cannot be classified according to the syntactic phrase they belong to, as this information is often irrelevant to the error produced, or cuts across two types of syntactic phrases. This is one example of the difficulties that are inherent to the creation of categories, which we discuss in 3.2.

3.1 Presentation of the classification

The following tables present the main categories of our classification, which are given in the headline. The second level of categories appears in the left-hand column, and takes the form of general linguistic phenomena (e.g. Determination), or categories (e.g. Adverb). The middle column is a list of the type of errors encountered in relation with these second level categories. For example, some of the errors found in the noun phrase are errors linked to adjectives, and among them some are due to an erroneous positioning of the adjective after the noun. The right-hand column gives one example of such erroneous segments (in italics), followed by the default correction.

NOUN PHRASE		
Adjective	Position of adjective w.r.t. noun	*The carrying of weapons is permitted in fifty states different.* The carrying of weapons is permitted in fifty different states.
	Order of adjectives in a complex construction	*European academic and industrial partners* Academic and industrial European partners
	Position of the adverb modifying an adjective (exceptional construction)	*A quite detailed analysis* Quite a detailed analysis
Determination	Choice of article	*A Merovingian necropolis was built on Ø exact site of the villa.* A Merovingian necropolis was built on the exact site of the villa.
NØN construction	Ungrammatical NØN construction	*The objects properties* The properties of the objects
	Abusive NØN stacking	*Security object granularity* The granularity of security objects
Morphology of the noun phrase	Determiner/noun agreement	*I didn't order this goods* I didn't order these goods
	Ungrammatical adjective/noun agreement	*News clothes* New clothes

Table 2.a

PREPOSITIONAL PHRASE		
Preposition	Choice of preposition according to the co-text	*They are exchanged and read on their electronic form* They are exchanged and read in their electronic form

Table 2.b

VERB PHRASE		
Order of elements following the verb	Separation of direct object and main verb	*Ontological domains include in our view objects, their properties and relations.* In our view, ontological domains include objects…
	Position of adverbial particle in phrasal verb	*It does not take into account context.* It does not take context into account.
Realization of verb-related lexical constraints	Choice of preposition	*These scores depend from the gold standard.* These scores depend on the gold standard
	Transitivity	*I'm waiting your answer.* I'm waiting for your answer.
Adverb	Position of adverb	*They exhibit nevertheless the dependency relationships observed in the source parse tree.* Nevertheless, they exhibit the dependency relationships…
	Use of adverbs of negation	*They are not only constrained to the author's point of view any more.* They are not constrained only to the author's point of view any more.
Aspect	Choice of aspect	*This summer, our association organizes a trip.* This summer, our association is organizing a trip.
Modal auxiliary	Choice of modal auxiliary	*It appears that patients who suffer from FRS would be unable to correctly monitor their actions.* It appears that patients who suffer from FRS are unable to…
Morphology of the verb phrase	Construction of compound tense	*You do not have takes action* You have not taken action

Table 2.c

CLAUSE AND SENTENCE		
Interrogative sentence	Construction of direct interrogative sentence	*It is possible to receive the parcel by the end of August?* Is it possible to receive the parcel by the end of August?
Subordinate clause	Construction of indirect	*It is necessary to know what is their role in the action*

	interrogative clause	*expressed by the predicate*. It is necessary to know what their role is in the action…
	Construction of non-finite clause	*They read annotations for evaluating them.* They read annotations to evaluate them.
Adjunct	Position of adjunct	*Goals and subgoals are most of the time realized by means of titles.* Most of the time, goals and subgoals are realized…
Comparative structure	Construction of comparative structure	*as many as possible of incorrect analyses* as many incorrect analyses as possible
Micro-planning	Un-idiomatic micro-planning	*August 15 in France, it is a holiday.* In France, August 15 is a holiday.
Morphology	Subject/verb agreement	*The first written mention of Issigeac date from 1008.* The first written mention of Issigeac dates from 1008.

Table 2.d

LEXICON and miscellaneous		
Lexical characteristics of term	Noun/verb confusion	*You will give my apologize to her.* You will give my apologies to her.
	Mass noun/count noun confusion	*A valuable information* valuable information
	Conjunction/adverb confusion	*Although, such an excess of mental effort should be reduced at all costs.* However, such an excess…
	Confusion between semantically close terms	[…] *to remind a document* to remember a document
Collocation and Idiomatic expression	Use of un-idiomatic expression	*Well cordially* Yours sincerely
Spelling	Spelling error	*A shedulle* a schedule
Punctuation	Choice of punctuation	*The purchase price will be validated by you and me, for the year*. The purchase price will be validated by you and me for the year.

Table 2.e

Let us point out that a number of errors are the conjunction of two or more problems: in that case, one may choose to classify this error according to only one of these, or to include it in all the categories concerned. For example, in the segment **They exhibit nevertheless the dependency relationships observed in the source parse tree*, the position of the adverb after the verb, which is in an error in itself, also results in a second type of error as the direct object is separated from the verb. This segment could therefore be included in the two corresponding categories.

3.2 Difficulties linked to the classification of errors

To begin with, the complexity of some errors may make their classification difficult. The erroneous segments that we find in emails very often contain errors which are be juxtaposed or embedded. For example, the following segment combines two morpho-syntactic errors (subjet/verb agreement and the construction of compound tenses), as well as an error concerning the choice of aspect: **Nobody have answer me*. In this case, it becomes quite tricky to find the more appropriate way to assign them to one single category.

The task of classifying errors is also problematic in itself. First of all, any classification system can be considered to be *ad hoc*, since it consists in the study of two definite languages, whose contact very often yields specific types of errors. For example, the categories of our system would undoubtedly be very different if we studied the English productions of native speakers of Thai.

Another type of difficulty stems from our initial choice of exploratory corpora (Albert et ali., in press). The heterogeneous nature of the productions taken into consideration (i.e. emails, reports, articles, scientific articles) results in the discovery of heterogeneous errors. For example, emails contain a wealth of morphological and lexical errors, but few errors linked to the syntax of the clause or the sentence, as writers of this type of productions avoid complex sentences and formulations. On the other hand, scientific publications contain very few lexical and morphological errors, these errors being easily corrected through editing and the use of spell checkers. This difficulty might be overcome thanks to adjustments in our corpora, as we have already mentioned in Section 1. (?). Nevertheless, the classification system that we have developed so far, and which is based on syntactic phrases and general linguistic phenomena, ensures that errors from different types of productions can be fitted into the same main categories. Distinctions thus become apparent in second-level categories.

At this stage of our study, the main flaw of the system chosen is its one-dimensional nature, as it is based on one single aspect of errors (i.e. the syntactic phrase and linguistic phenomena they are related to), and does not allow for errors to be distinguished according to other and equally important criteria, such as whether they are morpho-syntactic or lexical errors.

Finally, as this system of classification constitutes one of the steps in the realization of a software, one of the difficulties to take into consideration is the degree of granularity to give to second-level and third-level categories. If detailed categories enable the precise description of errors, they might be an obstacle to the implementation of results. Our objective is therefore to strike a balance between accuracy and usability.

3.3 The "ideal" classification system

The solution to some of the difficulties mentioned above might be to create a multi-dimensional system which would be able to gather and compile general linguistic information as to the nature of errors with information concerning the syntactic phrase they belong to, as well as the surface phenomena they result in (i.e. omission, addition, wrong order, etc.). Moreover, a satisfactory classification system should also enable the retrieval of information as to the cognitive source of the error (e.g. transfer, overgeneralization, etc.), in order to facilitate the production of relevant and useful corrections. These are two possibilities that we will shortly be investigating.

4. Annotating errors

Let us now introduce the annotation schema that we have developed so far. Further work is needed to stabilize the attributes and values, and the system needs to be improved and tested before it can be used to annotate greater amounts of corpora (Albert et ali, 2009). This schema is an attempt to reflect, in a factual and declarative way, the different parameters that emerged when we ran an analysis of the cognitive strategies used by human correctors as they detect and correct errors. We use a standard XML formalism, enriched with attributes whose associated values have a degree of granularity that has been established by the human correctors in our research group. The structure of these attributes has been designed in order to allow them to be used in an argumentation framework.

The next two subsections present the different steps of this annotation system, from the detection of an error to its correction. In the third subsection, we introduce some of the

difficulties encountered by annotators, as well as the different improvements that we are considering.

4.1 Delimitation and characterization of the error

The group of words involved in the error is tagged <error-zone>. The zone is meant to be as minimal as possible. This tag has several attributes, which are presented in the following table:

comprehension	from 0 to 4 (0 being worse), indicates if the segment is understandable in spite of the error
ungrammaticality	from 0 to 2, indicates how ungrammatical the error is
categ	gives the nature of the error (lexical, morphological, syntactic, stylistic)
source	gives the alleged cognitive source of the error (transfer, overgeneralization, etc.)

Table 3.a

4.2 Delimitation and characterization of the correction

The text fragment involved in the correction is tagged by <correction-zone>. It is equal to or larger than the error-zone.

Each correction is characterized by a tag <correction> and associated attributes. The attributes that are positively oriented are underlined in the following table:

surface	size of the text segment affected by the correction (minimal, average, maximal)
grammar	indicates, whenever appropriate, if the correction proposed is the standard one as suggested by grammar rules (by-default, alternative, unlikely)
meaning	indicates if the meaning has been altered (yes, somewhat, no)
var-size	is an integer that indicates the increase/decrease in number of words of the correction w.r.t. the original fragment
change	indicates if the changes in the correction are lexical, morphological, syntactic or stylistic
comp	indicates if the proposed correction is a text fragment which is easy to understand or not (yes, average, no)
fix	indicates, when mentioned, that the error is very specific to that string of words and that the correction is idiosyncratic and cannot be extended to any other such structure (yes, no)
qualif	indicates the certainty level of the annotator as to the quality of the correction proposed (high, average, low)
correct	gives the correction

Table 3.b

4.3 Example

Following is an example of annotation in the case where two corrections are possible:

Erroneous segment: "Although, such an excess of mental effort should be reduced at all costs"

<correction-zone>
<error-zone>
<comprehension="2" agrammaticality="0" categ="syntax/lexicon" source="calque">
Although, such an excess of mental effort should be reduced at all costs
<correction>
<surface="minimal" grammar="by-default" meaning="yes" var-size="0" change="lexical" comp="yes" fix="no" qualif="high" correct="However, such an excess of mental effort should be reduced at all costs">
</correction>
<surface="maximal" grammar="alternative" meaning="no" var-size="+?" change="syntactic" comp="yes" fix="yes" qualif="low" correct="Although X, such an excess of mental effort should be reduced at all costs">
</error-zone>
</correction-zone>

This example gives us the opportunity to mention the importance of interacting with the user in order to retrieve some of the information which is necessary to complete the correction (here, in the case of the second correction: "Although X, etc."). Moreover, we can see here how information concerning the source of the error, e.g. transfer, overgeneralization, wrong rule, etc., is an important asset in the proposition of relevant corrections: in this example, knowledge of the differences between French and English and of common mistakes of French speakers writing in English leads us to interpret the error as coming from the confusion of "although", which is a conjunction, with a contrastive adverb such as "however". A framework of argumentation, completed with decision theory, will be used in order to give advice to the user as to the most appropriate correction.

4.4 Difficulties

The main difficulty associated with this system is that it requires annotators with a high level of expertise in several domains. Preferably, these should be linguists who are highly competent in English, but also have a good understanding of French, since they should be able to recognize the presence of transfers from French to English. Ideally, these annotators should also be familiar with language teaching, and have some experience in the detection and correction of errors produced by non-native speakers of a language. These requirements make it quite difficult to recruit suitable annotators willing to conduct a complex annotation task.

Feedback from annotators revealed that they had difficulties in delimitating correction- and error-zones. For this reason we intend to design and issue more detailed guidelines as to what should be included in these zones. This will also be done in the case of relative and vague values such as "minimal, average, maximal". The definitions of attributes also need to be improved, as unstable definitions can cause inconsistencies in the results of different annotators. On the whole, annotators find it easier to perform the second part of the annotation, i.e. the delimitation and characterization of their own corrections, than the first part which deals with the delimitation and characterization of the errors encountered. The attribute *qualif* will therefore also be included in the attributes of <error-zone>.

Moreover, as we improve and stabilize our classification system, we will be able to include information as to the precise category of errors into the annotation schema. Research on

the cognitive source of errors (Garnier & Saint-Dizier, 2009) should also enable us to issue more rigorous guidelines concerning the annotation of the source of errors.

5. Conclusion and perspectives

In this paper, we introduced the preliminary steps of our project to create an intelligent error correction tool designed to improve the competence of French speakers writing in English. Our exploratory corpus was constituted according to several parameters, i.e. the diversity of authors, of fields of document production and of levels of control, and is composed of representative productions including scientific publications, web pages, professional and private emails. We intend to investigate the relevance of the use of already-existing learner corpora in our project, and the possibility of creating tailored corpora of that type.

We presented the system used to classify the errors found in the corpus. This system relies on linguistic criteria, i.e. the syntactic phrase the erroneous segment belongs to. The system also includes two other embedded levels of classification, so as to describe errors in more detail. In order to solve some of the difficulties encountered while using this classification system, we intend to create a multi-dimensional system which would enable the description of errors at several levels according to several dimensions. This new system would therefore include the syntactic information previously mentioned, and state whether the error is morpho-syntactic, lexical, or even stylistic. It could also give information as to the alleged cognitive source of the error, as this type of information is needed in order to provide relevant and efficient corrections.

Errors are annotated according to a system using XML attributes and their associated values. This system has been drafted in order to reflect the different parameters taken into consideration by human correctors as they detect and correct errors. It includes the steps of error delimitation and characterization, and delimitation and characterization of the correction(s) proposed. It is still at the stage of stabilization and requires modifications and improvements before it can be used on a larger scale. Attributes and their values are designed so as to enable the use of a framework of argumentation. Together with decision theory, this will be used to evaluate corrections and propose the most appropriate ones to the user. In some cases where some information is irretrievable, interaction with the user will be necessary to complete corrections. This system will also be used to draft rules for automatic correction. These are produced on the basis of induction from annotated examples. This machine-learning strategy, however rudimentary, is well adapted to our objective. The main difficulty will be to strike a balance between a proper level of abstraction and the treatment of particularities and exceptions. Focusing on pairs of languages introduces sets a relative limit to the types of errors that can be found, and thus enhances the feasibility of this task. These errors are basically inspired, or direct copies of well-formed structures of the source language. The innovative aspect of this approach is that it aims at describing the structure and particularities of errors, by means of a "grammar" of errors. It follows that detecting errors thanks to this method, rather than scan texts for any grammatical deviancies is much simpler and reliable.

References insuffisant

Albert, C., L. Buscail, M. Garnier, A. Rykner, & P. Saint-Dizier (2009). "Annotating language errors in texts : investigating argumentation and decision schemas". Proceedings of the ACL-LAWIII workshop, .

Albert, C., M. Garnier, A. Rykner, & P. Saint-Dizier (in press). "Eléments de stratégie de correction automatique de textes : le cas des francophones s'exprimant en anglais". Québec: Presses universitaires du Québec.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Garnier, M. & P. Saint-Dizier (2009). "An Analysis of the Calque Phenomena Based on Comparable Corpora". Proceedings of the ACL-BUCC workshop, 19-23.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot (2009). *International Corpus of Learner English v.2*. Louvain La Neuve: Presses Universitaires de Louvain.