An Analysis of the Calque Phenomena Based on Comparable Corpora

Marie Garnier
CAS, Université Toulouse le Mirail
31000 Toulouse France
mhl.garnier@gmail.com

Patrick Saint-Dizier
IRIT-CNRS, 118, route de Narbonne,
31062 Toulouse France
stdizier@irit.fr

Abstract

In this short paper we show how Comparable corpora can be constructed in order to analyze the notion of 'calque'. We then investigate the way comparable corpora contribute to a better linguistic analysis of the calque effect and how it can help improve error correction for non-native language productions.

1 Aims and Situation

Non-native speakers of a language (called the target language) producing documents in that language (e.g. French authors like us writing in English) often encounter lexical, grammatical and stylistic difficulties that make their texts difficult to understand. As a result, the professionalism and the credibility of these texts is often affected. Our main aim is to develop procedures for the correction of those errors which cannot (and will not in the near future) be treated by the most advanced text processing systems such as those proposed in the Office Suite, OpenOffice and the like, or advanced writing assistance tools like Antidote. In contrast with tutoring systems, we want to leave decisions as to the proper corrections up to the writer, providing him/her with arguments for and against a given correction in case several corrections are possible.

To achieve these aims we need to produce a model of the cognitive strategies deployed by human experts (e.g. translators correcting texts, teachers) when they detect and correct errors. Our observations show that it is not a simple and straightforward strategy, but that error diagnosis and corrections are often based on a complex analytical and decisional process.

Most errors result from a lack of knowledge of the target language. A very frequent strategy for authors is to imitate the constructions of their native language so that the production resembles standard terms and constructions of the target language. This approach based on analogy is called a *calque* when surface forms are taken into consideration (Hammadou, 2000), (Vinay et al. 1963). The errors produced in this context may be quite complex to characterize, and they are often difficult to understand. When attempting to correct these errors, we find it interesting to have access to some of the characteristics of the native language of the author so that a kind of 'retro-analysis' of the error can be carried out. This would allow a much better rate of successful corrections, even on apparently complex errors involving long segments of words in a sentence.

Works on the correction of grammatical errors made by human authors (e.g. Writer's v. 8.2) have recently started to appear. These systems do not propose any explicit analysis of the errors nor do they help the user to understand them. The approach presented here, which is still preliminary, is an attempt to include some didactic aspects into the correction by explaining to the user the nature of her/his errors, whether grammatical or stylistic, while weighing the pros and cons of a correction, via argumentation and decision theories (Boutiler et ali. 1999), (Amgoud et ali. 2008). Persuasion aspects are also important within the didactical perspective (e.g. Persuation Technology symposiums), (Prakken 2006). Finally, the calque (direct copy) effect has been studied in the didactics of language learning, but has never received much attention in the framework of error correction, where a precise analysis of its facets needs to be conducted.

In this short document we present the premises of an approach to correcting complex grammatical and lexical errors based on an analysis of the calque effect. Calque effects cannot easily be reduced to the violation of a few grammar rules of the target language: they need an analysis of their

own. For that purpose, we introduce several ways of constructing and annotating the forms calque effects can take in source and target language in bilingual corpora. These corpora are both relatively parallel, but also relatively comparable in the sense that they convey the same information even though the syntax is incorrect. From these annotations, different strategies can then be deployed to develop correction rules. The languages considered here are French, Spanish and English, which have quite rigid and comparable structures. We are investigating two other languages: Bengali and Thai, which have a very different structure (the former has a strong case structure and some free phrase order, the latter has a lot of optional forms and functions with a strong influence from context). Besides correcting errors, the goal is to make an analysis of the importance of the calque effect and its facets over various language pairs.

2 Constructing comparable corpora

2.1 General parameters of the corpora

The documents used to construct the corpora range from spontaneous short productions, with little control and proofreading, such as emails or posts on forums, wiki texts, personal web pages, to highly controlled documents such as publications or professional reports. Within each of these types, we also observed variation in the control of the quality of the writing. For example, emails sent to friends are less controlled than those produced in a professional environment, and even in this latter framework, messages sent to hierarchy or to foreign colleagues receive more attention than those sent to close colleagues. Besides the level of control, other parameters, such as target audience, are taken into consideration. Therefore, the different corpora we have collected form a continuum over several parameters (control, orality, audience, language level of the writer, etc.); they allow us to observe a large variety of language productions.

The analysis of errors has been carried out by a number of linguists which are either bilingual or with a good expertise of the target language. For each document, either a bilingual expert or two linguists which are respectively native speakers of the source language and target language were involved in the analysis, in order to guarantee a correct apprehension of the calque effect, together with a correct analysis of the idiosyncrasies and the difficulties of each language in the pair.

Calque effects cover a large range of phenomena. Here are three major situations, for the purpose of illustration:
(1) Lexical calque: occurs when a form which is specific to the source language is used; this is particularly frequent for prepositions introducing verb objects: *Our team participated to this project* where in should be used instead of to.
(2) Position calque: occurs when a word or a construction is misplaced. For example, in French the adverb is often positioned after the main verb whereas in English it must not appear between the verb and its object: *I dine regularly at the restaurant* should be *I regularly dine* ....
(3) Temporal calque: occurs for temporal sequences concerning the grammatical tenses of verbs in related clauses or sentences: *When I will get a job, I will buy a house* the future in French is translated into English by the present tense: *When I get a job*.

2.2 Scenarios for developing corpora

In (Albert et al. 2009), we present the different categories of errors encountered in the different types of documents we have studied, and the way they are annotated. These categories differ substantially according to text type. The approach presented below is based on this analysis.

In our effort to construct a corpus, we cannot use documents with several translations, such as notices or manuals written in several languages since we do not know how and by whom (human or machine) the translations have been done. In what follows, we present the two scenarios that seem to be the most relevant ones for our analysis.

A first scenario in constructing comparable corpora is simply to consider texts written by foreign authors, to manually detect errors (those complex errors not handled by text editors) and to propose a correction. Beside the correction, a translation of the alleged source text (what the author would have produced in his own language) is given. This study was carried out for the following pairs: French to English, French to Spanish and Spanish to English. So far, about 200 pages of textual document have been analyzed and tagged. The result is a corpus where the erroneous text segments are associated with a triple:
(1) the original erroneous segment, with the error category,

(2) the correction in the target language (since there may exist several corrections, the by-default correction is given first, followed by other, less prototypical corrections),
(3) the most direct translation of this segment into the author's native language, possibly a few alternatives if they are frequent. This translation is produced by a native speaker of a source language.

We have 22 texts representing papers or reports, about 20 web pages and about 80 emails or blog posts. These are produced by 55 different French authors, over a few domains: computer science, linguistics, health, leisure and tourism. Balance over domains and authors has been enforced as much as possible.

Here is an example based on our annotation schemas, mentioning some relevant attributes:
.... <error-zone error-type="future">
When I will get </error-zone>
<correction errror-rev="present">
When I get </correction>
<transl calque="future"> Quand j'aurais </transl>.....

A second scenario we are developing is to take existing texts in the source language, with a potentially high risk of calque effects, which are representative of the types of productions advocated above and of increasing difficulty, and to ask quite a large and representative population of users to translate these texts. Emails need to be translated in a short period of time while more formal texts do not bear any time constraints, so that authors can revise them at will. We then have a corpus which can be used to study how the calque effect functions and how it can optimally be used in automatic error correction.

In this latter scenario, important features are as follows:
Corpus: we built a set of short corpora (8 corpora), so that the task for each translator is not too long. Each corpus is about 5 pages long. It contains 2 pages of emails, some really informal and others more formal, 1 page in the style of a web page and 2 pages of more formal document (report, procedure, letter, etc.). Those texts are either real texts or texts we have slightly adapted in order to increase the potential number of calque effects.
Translators: we use a large population of translators (about 70), where the language competence is the major parameter. Age and profession are also noted, but seem to be less important. Each corpus is translated by 8 to 10 translators with different competences, so that we have a better understanding of the forms calques may take. Comptence is measured retroactively via the quality of their translations. For emails, translators are instructed to follow the provided text, possibly via some personal variation if they do not feel comfortable with the text. The goal is to improve naturalness (probably also in a later stage to study the forms of variations).
Protocol: in terms of timing, translators are asked to translate emails in a very short time span, which varies depending on the ability of the translator; conversely, they have as much time as needed for the other documents, which can be proofread over several days, as in real situations. No dictionary or online grammar is allowed.

3 Analysing the facets of the calque effect

Let us now briefly present how these corpora allow us to have a better linguistic analysis of the calque effect and how this analysis can help us improve error correction.

The first level of analysis is the evaluation of the importance of a calque error per category and subcategory. For the pair French to English, we are studying:

- lexical calques, among which: incorrect preposition, incorrect verb structure (transitive vs. intransitive uses), argument divergences (as for the verb to miss),
- lexical choice calques which account for forms used in English, which are close to French forms, but with different meanings *I passed an exam this morning* should be: *I took an exam this morning*, .
- structural calques, which account for syntactic structures constructed by analogy from French. In this category fall constructions such as the incorrect adverb position or the position of quite: *a quite difficult exercise* which must be *quite a difficult exercise*
- A few basic style calques, with in particular the problem of temporal sequence.

In terms of frequency, here are some examples of results related to calque effects, obtained from a partial analysis realized so far on 1200 lines of emails produced by about 35 different authors, for the pair French to English. Note that, in average,

emails have one error per line.
Lexical calques: incorrect lexical choice of preposition: 62, determiner: 30, adverbs: 12, modals: 26, incorrect idiomatic expression: 70.
Grammatical calques: incorrect position of adverbs: 38, adjectives: 7; argument omissions: 52, incorrect passive forms: 8.
Style: incorrect temporal sequences: 26, aspect: 20, punctuation: 76.

Alongside an evaluation of the distribution and frequency of the different categories of calque, in conjunction with the parameters considered in the corpus constitution (in particular foreign language level and type of document), we can analyze the evolution of the calque effect: when (i.e. at what language competence stage) and how they emerge, expand, and disappear. Another question is the analysis of the level of genericity of calques: some may be individual, related to the way a certain individual has experienced learning a foreign language, whereas some may be widespread among a certain linguistic population. Examining different document types is also interesting. It shows the performance of a subject when he must write hastily, with little control, in contrast with highly controlled productions. This allows us to analyze what remanence level of calques appear when the subject does not have the time to proofread his text, as opposed to those which are still present when he has time to proofread it. This also betrays a possible error hierarchy in the subject's mind, since the subject will be tempted to first correct the errors he thinks are the most important.

It is also interesting to take into consideration corpora over several language pairs, and in particular to contrast the French to English and Spanish to English pairs. Although French and Spanish are in the same language family, the calque effects observed are quite different. This is not surprising for a number of lexical calques, but more interesting for grammatical calques. For example, the grammar of pronouns and reflexives is quite different in Spanish, leading to forms such as *David is me*, a calque of *David soy yo*.

Finally, if we consider the two scenarios above, where the first one is probably a direct production in English, whereas the latter is a production via an explicit translation, it becomes clear that they require a different kind of effort. It is thus interesting to compare the frequency of the different calque categories encountered and their distribution over subjects. The translation from an explicit source is probably more constraining in terms of form and contents than text produced directly (or almost) in English. This is under investigation.

4 Perspectives

The work presented here is essentially the premises of a detailed analysis of the calque effect and, working on a language pair basis, on how this analysis can be used to substantially improve the performances of the correction for non trivial lexical and grammatical errors that current text editors cannot detect and correct. We have shown how corpora have been built. So far, they are quite small, but sufficient to make a preliminary and indicative analysis of the problems, and to suggest directions for research. These corpora are also too small to be used in any kind of statistical machine learning procedure to automatically correct errors.

Our goal is thus to propose some elements of a strategy for didacticians teaching foreign languages so that students can improve their performance, based on the knowledge of these effects.

References

Albert, C., Garnier, M., Rykner, A., Saint-Dizier, P., Analyzing a corpus of documents produced by French writers in English: annotating lexical, grammatical and stylistic errors and their distribution, Corpus Linguistics conference, Liverpool, 2009.

Amgoud, L., Dimopoulos, Y., Moraitis, P., Making decisions through preference-based argumentation. In Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR08), AAAI Press, 2008.

Boutilier, C., Dean, T., Hanks, S., Decision-theoretic planning: Structural assumptions and computational leverage. Journal of Artificial Intelligence Research, 11:194, 1999.

Chuquet, H., Paillard, M., Approche Linguistique des Problmes de Traduction, Paris, Ophrys, 1989.

Hammadou, J., The Impact of Analogy and Content Knowledge on Reading Comprehension: What Helps, What Hurts, ERIC, 2000.

Prakken, H., Formal systems for persuasion dialogue, Knowledge Engineering Review, 21:163188, 2006.

Vinay, Jean-Paul and Darbelnay, Jean: Stylistique Comparée du Francais et de l'Anglais, Paris, Didier, 1963.