A CORPUS-DRIVEN CONTRASTIVE PROJECT: CHARACTERISATION IN ENGLISH AND SPANISH

Noelia Ramón García
University of León

1. INTRODUCTION

The University of León has been engaged for several years now in a wide-reaching project on corpus-driven Contrastive Analysis (henceforth CA) involving two languages: English and Spanish. There are several communicative areas on which contrastive work is currently being done. These areas are chosen among those semantic functions whose representation in English grammar is particularly conflictive when translating into Spanish. The project is aimed at the setting up of manuals and other resources dealing with English-Spanish contrastive studies focused on corpus-driven data. The possible applicability of this project are mainly directed towards the improvement of Translator Training and Translation Practice, as well as to the development of Foreign Language Teaching (FLT) in Spain.

Due to the non-existence of a sufficiently representative English-Spanish translation corpus, the research project is based on two monolingual corpora: Cobuild/Bank of English and the Spanish CREA (Corpus de Referencia del Español Actual). Both are corpora of the general language in use, and they provide essential information for the development of our project.

The present paper will deal with how the planning for this particular project on characterisation in English and Spanish is designed and set up. There are several stages that must always be fulfilled in all the areas the study is going to cover. This is a proposal of the outline the final project will have.

2. THEORETICAL FRAMEWORK FOR CA

Firstly, it is necessary to choose the particular linguistic area that is going to be contrasted in the two different languages considered here. That particular linguistic area is usually a conflictive one, one that may cause problems when learning English or Spanish as a foreign language, or when translating from one into the other. The linguistic aspect that is going to be studied should thus present certain differences or a different scope in both languages. If there were no difference at all, no CA would be necessary. In this case, we have chosen the semantic function of characterisation. The reason for this choice is the fact that the frequent chains of two, three and more adjectives in front of a noun in English constitute an important area of difference between these two languages. Spanish is a language that does not admit these chains, and it tends to position adjectives after the nouns they refer to, and not in front of them. Similarly, the differences in the use of adverbs for describing verbal actions are another problem we are concerned with here.

Once we have chosen the part of the language we are going to talk about, it is necessary to determine the steps that are going to be followed in the research project. "Executing a CA involves two steps: description and comparison; and the steps are taken in that order." (James 1980: 63). Any CA presents, thus, these two stages in its development. With reference to the stage of description, it is important to point out that before we start dealing with any pair of languages to establish a CA, it is absolutely necessary to choose one common linguistic model as a frame of reference for the descriptive analysis previous to the comparison.

2.1. The appropriateness of Systemic-Functional Grammar for CA

Among all the linguistic theories that are sufficiently developed to be susceptible of being applied to CA, the one chosen for this general project is the Systemic-Functional Grammar proposed by M.A.K. Halliday. In recent times, several authors have been proposing the appropriateness of this model for the use in CA.

There are important reasons why this is the linguistic frame chosen for our project. Halliday's approach is termed *functional* because of the claim that the main characteristic of language is precisely its ability to convey meaning. Halliday does not refer to syntactic functions, or grammatical categories, but to semantic functions. The starting point for all our different subprojects, including this one on characterisation, are precisely semantic functions, and their realisations in the grammar of two languages by means of the formal resources they possess. The most important thing here is meaning, for it is the only thing two equivalent utterances in two different languages may have in common. It is precisely this importance of meaning pointed out by Halliday, which makes this model so suitable for our purposes.

> Because structure is so to speak, on the surface of language, it can be played with to great effect; but because it is not arbitrary, this play contributes to the overall making of meaning.
> (Halliday 1994: 16)

More recently, A. Chesterman (1998) has proposed Contrastive Functional Analysis as a proper methodology on its own for investigation in CA.

> It starts from perceived similarities of meaning across two or more languages, and seeks to determine the various ways in which these similar or shared meanings are expressed in different languages. [...] The perspective is from meaning to form. (Chesterman 1998: 1)

This approach particularly suits our project for many reasons. The starting point is meaning, and we build up our project going from meaning to the formal resources that each of the two languages involved uses spontaneously to express that meaning.

3. CORPUS LINGUISTICS & TEXT TYPOLOGIES

Before starting with the description of the linguistic elements chosen for this project, it is necessary to underline a few theoretical aspects of the working tools that are going to be used to test the hypothesis proposed after the descriptive study, i.e. computerised corpora. Modern corpora show one important characteristic, which is their extreme usefulness for extracting practical results in the field of CA. The inclusion of text typologies under this same heading is due to their usefulness in a contrastive study of the particular characteristics our project presents.

3.1. The usefulness of computerised corpora in CA

The development of all aspects related to computers has been very fast over the last decades. There are now in the market great numbers of computer programmes, corpora and other utilities that can be very helpful to research in linguistics in general and in CA in particular. Basically, we can say that a corpus is a group of texts, taken from a variety of sources, stored in machine-readable form, and prepared in such a way that information can be extracted from them.

There are numerous advantages in using computerised corpora in CA. Firstly, the important amount of words modern corpora contain makes all the results obtained fairly accurate and reliable. The two big monolingual corpora that we are going to use have many million words: Cobuild/Bank of English approximately 400 million words, and CREA around 100 million words. Secondly, as the texts included are real texts from the real world produced spontaneously by native speakers of the language (in the case of monolingual corpora), all the results obtained will be more objective and realistic. On top of that, the fact that modern corpora are stored in machine-readable form and the recently developed computer programmes allow a great variety of searching possibilities within the corpus. It is not only possible to search single words or groups of particular words, or seeing them in a bigger context than a mere concordance line. As the two corpora employed here are both grammatically tagged in detail it is also possible to search particular grammatical categories, or any other tag included in the corpus. This is especially important for our project, for there is no small closed class of words that is going to be analysed in the two corpora, but different grammatical structures that all serve to the same purpose: expressing the semantic function of characterisation.

3.1.1. The choice of corpora: Cobuild/Bank of English & CREA

Due to the special characteristics of the research project that is being presented here, a project where the starting point is not one particular word or group of words, but a whole semantic function, the choice of the corpora that are going to be employed is a decisive one.

First of all, there is no already existing English/Spanish translation corpus containing a sufficient number of words and being representative enough of the general language that could fit our purposes. Besides this, even if there were such a translation corpus, the results would never be so accurate as the ones that may be obtained from two monolingual corpora. In a translation corpus, there is always one half of the text that was not originally written in that language, but which is merely a translation of the other half. This fact favours the existence of the so called *translationese*, expressions of the translated text that are not natural in the contexts in which they appear due to the fact that they were not produced spontaneously in that language, but as a translation from a different language.

Furthermore, we can point out that the two corpora chosen as the tools for this CA have a series of common characteristics which make them similar enough to base our research on them. Both the Cobuild/Bank of English and the CREA are huge corpora of the general language. They are both well known and widely used around the world, and they are the biggest general corpora in their respective languages. We can state that they also enjoy a similar status of importance in the academic world of language research. Both corpora include 10% of spoken language versus 90% of written texts. They are divided into a series of subcorpora following similar criteria. These subcorpora include oral texts, newspaper texts, academic writings, etc. all articulated according to the principle of the geographic origin of the texts: British versus American English, and European versus South American Spanish. This fact is important and can be very useful when it comes to restricting the search to certain subcorpora, for we could choose two very similar ones in the two corpora.

Finally, the fact that the search would have to be based primarily on grammatical categories and not on single words requires the use of highly developed and grammatically tagged corpora. These two corpora are grammatically tagged in a detailed way. Even though the semantic function of characterisation is mainly expressed by means of the formal resources of adjectives and adverbs in both languages, it is not only impossible, but also useless to track them all down in the two corpora one by one. We need the possibility of searching whole grammatical structures such as adjective+adjective+adjective+noun (JJ+JJ+JJ+NN in Cobuild/Bank of English, for example). Thus, the comparison of a whole list of formal structures expressing characterisation in both languages would provide enough information for a very detailed CA in this semantic field. That is the final aim of the present project.

3.2. Text typologies & CA

It sometimes turns out to be necessary, for the correct development of certain types of linguistic studies, either to establish a new text typology or to follow one that has already been proposed previously. One might argue that for a CA at the microstructural level such as this one text typologies need not be taken into account. Nevertheless, we are going to approach this aspect of linguistics here for two reasons.

Firstly, the topic of the research being characterisation, it might be useful to establish a typology including a particular text type that contains frequent descriptions of nouns (persons, animals or things) and of verbal actions or situations. It is true that any type of text may include some aspects of characterisation, for this semantic function is extremely common and has an enormous scope. However, some texts could be clearly marked in this aspect, thus presenting a much higher frequency of appearance of characterising structures.

The second reason why a text typology could be useful in this case is related to the huge scope of the semantic field of characterisation. Every time we formulate a message, in the oral or in the written form, we are communicating to somebody information about something. In a wide sense, practically everything we can say is characterising in one respect or another. If we could limit our research exclusively to those texts that belong to the particular text type of so called *characterising* texts, we would reduce the scope of our investigation to a smaller and easier manageable amount of texts.

3.2.1. The choice of a text typology

Among all the text typologies proposed by different authors over the last thirty years, there is one particular tendency that suits our purposes better than the others. R. Beaugrande and W. Dressler (1981) distinguished between two big text types, depending on whether the function of the text was a merely descriptive one, or whether its function was rather aimed at producing a change of any type in the situation presented.

> If the dominant function of a text is to provide a reasonably unmediated account of the situation model, SITUATION MONITORING is being performed. If the dominant function is to guide the situation in a manner favourable to the text producer's goals, SITUATION MANAGEMENT is being carried out. (Beaugrande & Dressler 1981: 162)

This text typology is fairly simple and concise, which is a clear advantage in this case. Other authors had previously distinguished seven, eight or even many more text types, whereas a basic typology is clearly easier to deal with and better suited for many linguistic purposes. Why make things complicated if they can be as simple as this?

B. Hatim (1997) revised this two-fold typology and proposed a division into three text types: *expository texts*, which comment on a certain idea in a relatively objective way; *argumentative texts*, which reflect a particular attitude with respect to the topic in the text; and *instructive texts*, which convey certain instructions to the reader or listener of the text. Each of these types presents further subdivisions.

- Exposition: conceptual exposition, narration, description.
- Argumentation: through-argumentation (thesis cited to be argued through), counter-argumentation (thesis cited to be opposed).
- Instruction: without option, e.g. 'contracts, treaties', etc., with option, e.g. 'advertising'. (Hatim 1997: 39)

Using this text typology Hatim is able to define with great accuracy the different text types in typologically very different languages. This is of particular interest for his own purposes, for his studies take into account the comparison of English and Arabic text types. It is also the typology that suits best our purposes in this case, since it establishes a clear distinction between exclusively *expository texts* and any other kind of texts. Apart from this, one of the subtypes within the text type *exposition* is actually called *description* and is defined as "dealing with 'objects' or 'situations'" (Hatim 1997: 37). Using this typology as a basis would be very helpful in several ways. To begin with, it would provide us with a text type framework we can easily apply to both languages English and Spanish. It could also allow the restriction of the field of study to certain only *expository* or even only *descriptive* texts, thus making the task much easier to cope with, and centred in the particular text type where the structures we are looking for are more common.

4. WORKING PROCEDURE

Once all these theoretical issues are dealt with, once the linguistic model of description is chosen, and once the testing tools are presented, it is necessary to prepare an outline of the different parts this research project will consist of. The following headings constitute one possible approach to a CA of the characteristics we have mentioned above. It is important to follow these steps in the order proposed for the results found in every stage are going to be an important part of the material that is going to be used for the next steps.

4.1. Delimitation of the field

What is characterisation? Or better, what do we understand by characterisation? Characterisation is an ambiguous word that may involve in a wide sense any aspect of language that is susceptible to be used to say something about something else. A delimitation of the field is therefore unavoidable, especially in this case, for we cannot cope with the whole language as our field of contrast.

For our purposes, we will consider that characterisation is the semantic function expressed by any formal resource a language possesses to add qualitative information to the context. We will thus deal with the resources expressing quality, i.e. those resources that add - to nouns or verbs - certain information, that cannot be directly deduced by the context or by the lexical meaning of those linguistic elements. This limits our field of study to certain formal entities such as adjectives and adverbs, and more complex structures with equivalent meanings, such as subordinate clauses.

4.2. Systematic linguistic description

A systematic description of the chosen resources in English, as well as in Spanish, is necessary at this point to allow the comparison between the two languages. This is a hard task, for it is necessary to unify the criteria for the description of adjectives and adverbs in both languages. Common classifications, typologies and categorisations should be produced so that the two systems can be superposed and compared later on. This task requires an important effort on the part of the researcher, who will have to consult a great number of specialised publications to find the linguistic descriptions that suit best his purpose, or to build up a new way of approaching the topic taking into account the particular needs of the project.

4.3. Listing of formal entries

The next part of the project will be the listing of all the formal resources that are going to be taken into account. This is an important decision to take, for the field to cover is still very big, even after the previous delimitation. As we will not search every single adjective or adverb one by one, it is necessary to set up a list of the structures that are going to be searched on the basis of their grammatical category. For example, in the case of English, some of the elements of this list would be: adjective+noun, adjective+adjective+noun, adjective+adjective+adjective+noun, etc. And in the case of the Spanish language, we could search strings such as: adjective+noun, adjective+noun+adjective, noun+adjective, noun+adjective+adjective, etc.

As the scope is still huge, one reasonable option to continue this research and obtain representative results is to restrict the search to certain subcorpora, which must, of course, be present in both corpora, for example, the written media. Another option for reducing the scope could be the restriction to very peculiar formal entries in the general corpus, such as for example only the strings of more than two adjectives in front of one noun in English, leaving out other simpler structures. These chains of adjectives in English have already been mentioned above as one of the main translation problems into Spanish. Of course, these two options may be combined to simplify the search further.

4.4. Actual search in the corpora

Once there is a closed list of the structures to be searched in both corpora, the actual contrastive research may start. It is necessary to check the formal strings chosen in the corpora to obtain representative results. This task has to be made by the researcher himself or herself, by looking for concordances, analysing them carefully one by one manually, and making percentages of frequency for both languages. There is no machine that has been invented that may substitute the researcher when it comes to doing the actual research work. Computers provide a great help, but it is still the researcher who has to be able to extract conclusions from the information he has access to.

For our purposes, no CA can be adirectional, but has to be necessarily bidirectional. Our way of proceeding is the following: we search a particular structure in the English corpus, then look for the equivalent one, if there is one, in the Spanish corpus, and then go back again to the English corpus to confirm the accuracy of the results found. Nevertheless, there is one important problem we have to face, and that is related to what we understand as *equivalent*. It is the main problem every CA has had to face over the years: there are no exactly equivalent structures in any two languages, or at least that fact is not common, not even in typologically closely related languages. How can we ever be sure that we are searching similar structures if they do not cover exactly the same scope? The explanation of this problem will be commented on in the next section.

4.5. The paradox: semantic functions versus formal entries

All the problems that we have encountered while planning this project are due to one single and very simple paradox: even though our starting point is a semantic function, the one of characterisation, the only tools we have to work with are formal entries in two different corpora. Due to the fact that corpora are not semantically, but grammatically tagged, the only possibility for us is the search of formal strings or grammatical categories that are known to express that semantic function of characterisation.

Some authors have argued that the expression of certain meanings, of certain semantic functions, requires and even determines the use of particular formal structures. This is the so-called syntax-semantics interface. If we could exactly determine which formal structures are used to express which meanings, all our problems would be solved at once. Nevertheless, this is only a utopian vision of reality, which is much more complex than this. Making formal entries conform to semantics is a very hard task indeed. On the other hand, these formal entries are the only way to gain access to more information about the different ways we choose to convey certain meanings. And nowadays, a computerised corpus is a necessary and reliable tool, which yields valid and representative results in CA. We have thus to accommodate to the possible problems caused by its use in linguistic research, and try to look for solutions to those problems.

5. CONCLUSION

In the present paper, we have tried to make a proposal of how a CA research project based on a semantic function and on corpus-driven data can be carried out. We have outlined the theoretical basis that will have to be set up first and the different stages in the working procedure that will have to be covered later on. As we have said above, the use of Halliday's Systemic-Functional approach to language allows for a contrastive study of the characteristics of the project described here. The use of meaning as the only *tertium comparationis* possible here this case makes the future results of this project seem very promising, not least because very little research has been carried out up to the moment taking English and Spanish as the languages to be contrasted.

With reference to the practical applications of the results of this study and the other studies within the same bigger corpus-driven contrastive project, it is important to point out that they will be of great importance in the Spanish-speaking world. They will finally provide Spanish teachers of English as a foreign language with important contrastive material on which to base their courses. However, the field where the results are expected to have a greater resonance is translation. English is by far the language from which more texts are translated into Spanish every day. Even though the quality of translations is constantly improving, the information that may be provided by this contrastive project is hopefully going to have a very important repercussion on the acceptability of the Spanish translations from English.

REFERENCES

AIJMER, K. & ALTENBERG, B. (eds.) 1991: *English Corpus Linguistics*. London: Longman.

AIJMER, K.; ALTENBERG, B. & JOHANSSON, M. (eds.) 1996: *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund: Lund University Press.

BEAUGRANDE, R. & DRESSLER, W. 1981: *Introduction to Textlinguistics*. London: Longman.

BIBER, D. et al. 1998: *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

CHESTERMAN, A. 1998: *Contrastive Functional Analysis*. Amsterdam: John Benjamins Publishing Company.

HALLIDAY, M.A.K. 1994: *An Introduction to Functional Grammar*. Arnold. London.

HATIM, B. 1997: *Communication across Cultures. Translation Theory and Contrastive Text Linguistics*. Exeter: University of Exeter Press.

JAMES, C. 1980: *Contrastive Analysis*. London: Longman.

SPERBER, D. & WILSON, D. 1995: *Relevance. Communication and Cognition*. Oxford: Blackwells.

THOMAS, J. & SHORT, M. (eds.) 1996: *Using Corpora for Language Research*. London: Longman.