CORPUS-BASED CONTRASTIVE ANALYSIS AND THE TRANSLATION OF SPECIALIZED TEXTS

IAN A. WILLIAMS

Universidad de Cantabria

This paper presents an outline of, and some details from, an exploratory contrastive study of citations in medical research articles (RAs) based on a specially-designed corpus. The first section considers some theoretical implications of contrastive studies. The second deals with certain aspects of corpus design. The final part illustrates the approach with data from a lexical analysis of one of the citation categories.

1. CONTRASTIVE STUDIES

A complete contrastive study can be said to have three steps: description, juxtaposition and comparison. The first step is to define the phenomena to be compared; the second establishes the matching of equivalents across languages; and the final step is often considered the contrastive analysis proper, for example, the confrontation of source language (SL) and target language (TL) texts.

In the first step, two questions arise: *a*) whether the object of study is comparable across languages, and *b*) what means of comparison, or *tertium comparationis*, is to be used. In our case, citations in RAs fulfil certain communicative functions, which are an essential part of the RA and, therefore, universal. However, citations have diverse forms and these may be represented to different degrees across languages. Thus, the type of equivalence applied in the study was not formal (i.e. based only on semantics and syntax), but functional, taking in pragmatic considerations. Since citations have a communicative function, they represent a type of speech act and were classified according to a speech act model proposed by Oleksy:[1]

— citations with a finite reporting verb (e.g., *demostrar*) were labelled R+;

— citations with other markers, such as names of researchers (*Donaldson et al. evaluaron*), general nouns indicating authors or previous research (*investigadores*, *estudio*), or reporting verbs with a defining function (*desencadenante demostrado*), were classed as M+;

— citations with no marker in the body of the text and signalled only by a superindex number or parenthetically were M–.

As with other speech acts, there is no one-to-one relation between form and function, and any category can provide background knowledge, justify a new study, compare re-

[1] W. Oleksy: "Towards pragmatic contrastive analysis", in J. Fisiak (ed.): *Contrastive Linguistics: Prospects and Problems*, Berlin, Mouton, 1981.

Lengua y Cultura. WILLIAMS, Ian A. Corpus-based contrastive analysis and the tra...
Centro Virtual Cervantes

sults, or support a statement, hypothesis or recommendation. What remains to be seen is whether English and Spanish differ with respect to how often the functions are used and how they are realized.

The second step is often referred to as theoretical contrastive analysis, which establishes the basis for comparison in step 3 by analyzing the two languages independently. For a specialized text corpus, three main points must be borne in mind. First, the analysis is text-bound, meaning that any conclusions are restricted to the corpus studied or to similar texts. Second, the kind of equivalence established is statistical. Finally, the constituent texts should be restricted to the same register, the same topic, or the same literary genre so that comparability is guaranteed.[2] In the case under study, this was achieved by using concept of parallel texts, as defined by Neubert and Shreve:

> True parallel texts are not the results of previous translation. They are results of a process of parallel evolution. They spring from similar cultural needs to serve congruent interests in comparable situations.[3]

With independently-created texts, we can determine empirically whether congruent forms appear with the same or different frequency across languages, and we can establish the relative frequency with which different forms realize a particular function or occur in a specific section of the RA.

In step 3, when the translations are compared with the independent parallel texts, we determine whether the translations conform to the pattern of frequencies (statistical norm), or whether, and to what degree, they deviate from it; in other words, whether the translators have given priority to the SL or TL system, the criteria of adequacy and acceptability, respectively.[4] Krzeszowski[5] measures deviation in terms of "errors of abundance" and "errors of avoidance", but these errors are stylistic in nature because the concern is primarily frequency of use.

2. CORPUS CREATION

The considerations for creating both general and specialized corpora are similar, but specialized material is best kept separate from a general corpus.[6] The most important concerns are the size of the corpus and the size of the individual samples. Although a corpus should be as large as possible, a corpus based on subsets of 10 texts per genre has been found to give highly reliable results for linguistic features such as passives, prepositions, tense markers, etc.[7] However, these are high frequency features. When lower frequencies are expected, the number of samples must be increased. Individual samples of 1,000-words extracted from longer texts have also been shown to display the distribution

[2] T. P. Krzeszowski: "Quantitative contrastive analysis", *Studia Linguistica* (1981), 35, pp. 102-113.

[3] A. Neubert & G. M. Shreve: *Translation as Text,* Kent-Ohio, Kent UP, 1992, pp. 89.

[4] G. Toury: *In Search of a Theory of Translation,* Tel Aviv, Porter Institute for Poetics and Semiotics, 1980.

[5] T. P. Krzeszowski: "Quantitative contrastive analysis", *Studia Linguistica* (1981), 35, pp. 102-113.

[6] J. Sinclair: *Corpus, Concordance, Collocation,* Oxford, OUP, 1991.

[7] D. Biber & E. Finnegan: "On the exploitation of computerized corpora in variation studies", in K. Aijmer & B. Altenberg (eds.): *English Corpus Linguistics,* London: Longman, 1991, pp. 204-220.

Lengua y Cultura. WILLIAMS, Ian A. Corpus-based contrastive analysis and the tra...
Centro Virtual Cervantes

of features in a text reliably.[8] Nevertheless, most researchers recommend at least 2,000-word samples. The alternative is to use whole texts.

In the study of citations, each subcorpus consisted of 17 RAs. These generally vary in length from about 3,000 to 10,000 words. The total number of words included in the citations ranged from 13,000 to 15,500 in the four subcorpora. This was considered adequate for this exploratory study but a larger corpus is necessary for a more rigorous analysis.

The creation of this particular corpus (Fig. 1) was determined by certain special features. First, of the vast number of RAs published in English in the biomedical field, only a small proportion are published in Spanish. Such publications are not always easy to obtain from university hospital libraries; of a total of nine journals translated in the year selected, only six were available in Spanish. Second, statistical treatment requires random selection, but for the descriptive purposes of the study, I used a semi-random system, selecting the first RA from the January, July and December issues. Finally, the subcorpus of Spanish translations (TL texts) had to be created first. The SL texts were then located. Finally, parallel subcorpora in English and Spanish were constructed from journals covering the same disciplines as the translated texts: general medicine, surgery, pediatrics, and obstetrics and gynecology.

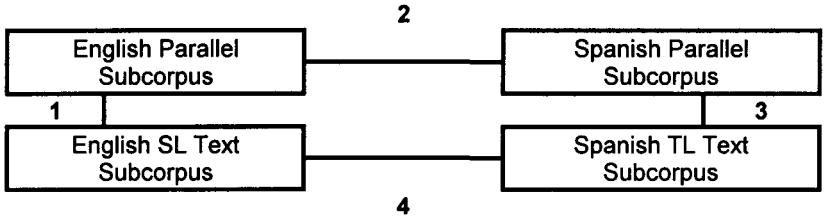

Figure 1: Four-way Corpus Design with Four Comparisons between the Subcorpora

This four-way design allowed four comparisons to be made. Comparison 1 assessed whether the set of SL texts was representative of the genre within the disciplines studied. Comparison 2 established the statistical norm for each language. Comparison 3 revealed whether the translators followed the TL norm or whether there were errors of abundance or avoidance. Finally, comparison 4 indicated whether the translators gave priority to the SL norm, i.e. adequacy as opposed to acceptability.

The study analyzed citations at various levels: discoursal, semantico-syntactic, grammatical and lexical. Within the R+ and M+ categories, a semantic-syntactic hierarchy was established as follows: human agent (specific name or generic noun), discourse or research element, referential element with passive verbs, and impersonal structures.

3. RESULTS

The four comparisons produced the following general results. The English parallel texts and the SL texts were similar in nearly all respects, and the SL texts were, therefore, considered representative. The English and Spanish parallel subcorpora showed similari-

[8] Ib., p. 212.

Lengua y Cultura. WILLIAMS, Ian A. Corpus-based contrastive analysis and the tra...
Centro Virtual Cervantes

ties in those features determined by the genre, such as the distribution of citations in the different sections, and the communicative functions of citations. However, there were marked differences at the semantico-syntactic, grammatical and lexical levels. For example, in Spanish discussions, the M+ category outnumbered the R+ category, whereas in English this was reversed. This difference was mainly due to two particular patterns.

1.a *Con esta metodología,* Rouvy et al. obtienen *una sensibilidad del 80% y una especificidad del 66%.*

1.b *Ballal et al. reported* sensitivities from 67 to 83 percent.

The Spanish participant has a research verb (M+) where English prefers a reporting verb (R+).

2.a Para algunos investigadores, *la reintervención será precoz.*

2.b Fine needle aspiration *has been advocated* recently as an alternative procedure.

In Spanish, the citation is introduced by an adjunct (M+), whereas English would probably use a reporting verb (R+). There were 13 adjuncts of this type in the Spanish sample but none were found in the two English subcorpora.

Comparison of the TL texts and the Spanish parallel texts revealed major differences wherever the Spanish and English parallel texts diverged. In addition, the TL and SL text frequencies differed only slightly. Thus, when translating citations in medical RAs, translators give priority to the SL system, and apply the principle of adequacy, rather than acceptability.

4. LEXICAL VERB ANALYSIS

The lexical verb analysis illustrates this point. Table 1 gives the frequency of finite reporting verbs in the 4 subcorpora. The first four verbs in the two left-hand columns occupy the same relative positions. They also account for over half of the total. A further eight verbs appear at least twice in each sample (*describe*, *demonstrate*, *note*, *know*, *confirm*, *advocate*, *accept*, *indicate*). The top twelve verbs now represent nearly 80 percent of the total.

Of the first four verbs in the right-hand column, *demostrar* and *encontrar* are semantically equivalent to *show* and *find*, *describir* covers some functions of *report*, but *considerar* and *suggest* have little in common. The four Spanish verbs have a lower frequency, making up only one third of the total, while the top twelve represent 60 percent. In addition, column 4 contains 50 verbs, whereas the two English subsamples only have 47 different verbs for twice as many instances. The figures so far support the notion that English prefers lexical consistency and repetition, whereas Spanish favours variety and synonymy.

A comparison of columns 3 and 4 reveals clear "errors of abundance": *sugerir*, *comunicar*, *observar*, *mostrar*, *informar* and *documentar* all rise considerably as translations of the four leading English verbs. The most notable "error of avoidance" is *considerar*, together with *confirmar*, *proponer*, *señalar* and *conocer*.

Centro Virtual Cervantes

Table 1
Reporting Verbs in the Different Subcorpora in Frequency Order

English Parallel Texts		English SL Texts		Spanish TL Texts		Spanish Parallel Texts	
report	22	report	30	demostrar	19	*considerar	13
show	19	show	22	sugerir	18	demostrar	13
suggest	15	suggest	18	comunicar	12	+describir	10
find	10	find	15	observar	11	encontrar	9
demonstrate	9	describe	7	describir	9	confirmar	8
note	8	recommend	7	mostrar	9	proponer	6
describe	6	confirm	4	recomendar	8	*+señalar	5
know	5	know	4	encontrar	6	+ comprobar	4
advocate	3	advocate	3	informar	5	conocer	4
believe	3	demonstrate	3	confirmar	4	hallar	4
accept	2	document	3	documentar	4	destacar	3
cite	2	note	3	saber	4	recomendar	3
concur	2	quote	3	comprobar	3	* sugerir	3
confirm	2	accept	2	considerar	3	+ comunicar	2
indicate	2	indicate	2	aceptar	2	concluir	2
postulate	2	regard	2	citar	2	+ constatar	2
state	2	advise	1	defender	2	establecer	2
agree	1	believe	1	determinar	2	*+hacer referencia	2
blame	1	calculate	1	indicar	2	insistir	2
conclude	1	conclude	1	juzgar	2	+ observar	2
consider	1	consider	1	poner de manifiesto	2	* opinar	2
delineate	1	discuss	1	afirmar	1	+ recoger	2
document	1	endorse	1	calcular	1	recordar	2
encourage	1	estimate	1	concluir	1	+ referir	2
estimate	1	implicate	1	creer	1	+ reflejar	2
expect	1	observe	1	discutir	1	saber	2
implicate	1	point out	1	especular	1	aceptar	1
predict	1	postulate	1	estimar	1	admitir	1
question	1	predict	1	hablar	1	afirmar	1
recommend	1	presume	1	implicar	1	+ aportar	1
review	1	reckon	1	notar	1	+ apreciar	1
say	1	record	1	postular	1	* apuntar	1
theorize	1	review	1	predecir	1	asumir	1
view	1	speak	1	publicar	1	+ cifrar	1
				registrar	1	comentar	1
				suponer	1	corroborar	1
				tener en cuenta	1	creer	1
						decir	1
						estar de acuerdo	1
						expresar	1
						imaginar	1
						implicar	1
						+ informar	1
						mostrar	1
						+ objetivar	1
						* pensar	1
						preconizar	1
						+ publicar	1
						+ registrar	1
						suponer	1
Total	131	Total	146	Total	145	Total	135

* verbs in the Spanish subcorpus of parallel texts within the semantic range of *suggest*
\+ verbs in the Spanish subcorpus of parallel texts within the semantic range of *report*

Lengua y Cultura. WILLIAMS, Ian A. Corpus-based contrastive analysis and the tra...
Centro Virtual Cervantes

The reporting verbs were also grouped according to semantic areas. Research processes that are also reporting verbs were classed as *objective* (e.g., *encontrar*) or *effect* (*demostrar*) verbs, the latter indicating that the citing author has been convinced by the evidence. Verbs expressing acts of speaking or writing were either **informative** (*describir*), *persuasive* (*insistir*), or expressed *doubt* (*implicar*). Finally, verbs expressing mental acts were classed as *cognitive* (*creer*). Table 2 gives the figures for these verb classes in the Spanish and English parallel texts in descending order.

Table 2
Distribution of Reporting Verb Classes (number and percent) in the English and Spanish Subcorpora in Descending Order

English Parallel Subcorpus		**Spanish Parallel Subcorpus**	
Informative	36 *(27.5)*	Cognitive	31 *(23.0)*
Effect	30 *(22.9)*	Informative	26 *(19.2)*
Doubt	25 *(19.1)*	Objective	25 *(18.5)*
Objective	18 *(13.7)*	Effect	25 *(18.5)*
Cognitive	16 *(12.2)*	Persuasive	24 *(17.8)*
Persuasive	6 *(4.6)*	Doubt	4 *(3.0)*
Total	**131** ***(100)***	**Total**	**135** ***(100)***

In English there is a stepwise descent between categories of about 4 to 5 percent, which reflects the order of the four main verbs (*report, show, suggest, find*). In contrast, the major class in Spanish was cognitive, doubling the English frequency. The next four classes were almost identical, with only a small proportion of verbs expressing doubt. Apart from cognitive verbs, the main differences are seen in the reversal of the doubt and persuasive classes.

The verbs *suggest* and *report* will now be considered as examples of two different problems for the translator. *Suggest* was invariably translated as *sugerir* in the TL texts, thus producing an "error of abundance". Some translators class *suggest* as a false friend:

> *Muchos autores en lengua inglesa usan hasta la saciedad este verbo. No es recomendable abusar también en castellano del verbo* sugerir.[9]

In the right-hand column of Table 1, the verbs marked with an asterisk were found in context to be within the semantic scope of *suggest*. Using the corpus data as a guide, I propose a number of strategies to avoid the overuse of *sugerir*. First, when a human agent is available, a cognitive verb (*considerar, opinar*) or a persuasive verb (*señalar*) can be used. Second, when there is no human agent and the subject is a discourse component, verbs with stronger implications can be used in an attenuated form: *viene(n) a confirmar, parece(n) indicar*. Moreover, since emphasis falls on the cause-effect relationship, in which data or findings serve as a stimulus for thought and deduction, this can be made explicit in the verb form: *hacer pensar, llevar a creer, deducirse, des-*

[9] F. A. Navarro & F. Hernández: "Palabras de traducción engañosa en el inglés médico", *Medicina Clínica* (1992), 99, p. 579.

Lengua y Cultura. WILLIAMS, Ian A. Corpus-based contrastive analysis and the tra...
Centro Virtual Cervantes

prenderse. Third, a human agent can be included in a fronted adjunct (*para/según unos autores*), but the main verb must then indicate the speculative nature of the statement through *poder*, or a lexical verb in the future or conditional. The citation thus changes category in translation (R+ to M+), in agreement with the above-mentioned pattern.

Close analysis of the 18 contexts in which *suggest* appeared in the SL texts indicated that translations could involve 11 different verbs (*considerar*, *pensar*, *confirmar*, *señalar*, and *sugerir* all twice; plus *afirmar*, *apuntar*, *creer*, *deducir*, *implicar* and *indicar* once each), with adjuncts being possible in two contexts. Such a solution would bring us much closer to the Spanish reader's expectations of stylistic variation and a more natural frequency.

The verb *report* presents different problems because it has no true cognate form in Spanish. In the SL texts, in addition to the 30 finite forms, *report* occurred a further 11 times in non-finite form, mainly as a past participle. Of the 41 forms, 4 were translated by other means and 37 by 11 different reporting verbs, notably by *comunicar* 15 times, many more than in the parallel texts. Apart from the frequency problem, use of *informar* requires a preposition and this can result in a very heavy style for a sentence element that simply adds author attitude to the propositional content.

3.a Intriguingly, Naeye (12) reported that...

3.b *Fue sorprendente el descubrimiento de Naeye (12),* que informó de que...

In four words, the English adds two aspects, interest and neutrality, to colour the proposition, whereas the Spanish takes ten words to achieve the same. A better solution might be: *Destaca la observación de Naeye de que...*, where interest and objectivity are reflected in a more natural style.

A second problem is collocation. *Report* combines with almost anything, but *comunicar*, *documentar*, *publicar* are more restricted in this respect. *Describir*, like English *describe*, collocates not only with techniques and procedures but also with other phenomena in the medical setting (e.g., *hemorragia*, *caso*), where English alternates *report* and *describe*.

As with *suggest*, the verbs in column 4 of Table 1 marked with a plus sign fell within the scope of *report* in the parallel corpus. The main strategy to avoid the problems identified is to focus not on the writing stage but on the research by using *observar*, *encontrar*, *hallar*, *comprobar*, or *constatar*. This may entail the addition of adjuncts to mark the citation: *en otros estudios, se ha comprobado que...*; or *observados por otros autores*, which may be the most natural translation of *previously reported*, since the Spanish past participle is best completed by the agent. Persuasive verbs like *señalar*, *destacar* can also be used in addition to the neutral *recoger*, *referir*, *reflejar*, *cifrar*, when the context indicates they are suitable choices. The final strategy is to use an adjunct (*para/según estos investigadores*) followed by a factual proposition with the verb in the present or the past.

For the 41 contexts in which *report* appeared, as many as 15 different reporting verbs could be used, the most frequent being *describir* (8), then *comunicar* and *observar* (4 each), *publicar* and *comprobar* (3), *informar*, *constatar*, *encontrar* and *hallar* (2), *señalar*, *destacar*, *reflejar*, *hacer referencia*, *recoger* and *cifrar* (once). Again, this distribution more closely reflects the expected frequency pattern.

Lengua y Cultura. WILLIAMS, Ian A. Corpus-based contrastive analysis and the tra...
Centro Virtual Cervantes

5. CONCLUSIONS

Corpus-based contrastive analysis offers a framework in which insights and intuitions about language patterns and translation can be tested empirically. The results of such contrastive analysis can provide the basis for descriptive studies, and for establishing translation strategies that can be applied consistently to specialized texts in an effort to achieve a more natural style.

Lengua y Cultura. WILLIAMS, Ian A. Corpus-based contrastive analysis and the tra...
Centro Virtual Cervantes