THE IMPLICATIONS OF A VARIATION CORPUS IN DOMAIN SPECIFIC ENGLISH

Mª Luisa Carrió Pastor
Universidad Politécnica de Valencia

Abstract:
English is widely used all around the world by native speakers (NS) and non-native speakers (NNS), as it is considered a lingua franca. The scientific community is conscious of this fact; as a consequence, contrastive studies about second language use have been increasingly attracting scholarly attention. In this article, we are going to refer to language variation as the different language production performed by NS of English and NNS of English. It can easily be noticed that writers of academic papers use some words or structures with different frequency in the same context. The objectives of this paper are to demonstrate that a corpus including the variations found in a standardised context as scientific articles can illustrate the parts of the sentence that are more sensible to variation and it can also make obvious the incidence of variation in a written text, evidencing the non-standardisation of language use. In order to fulfil these objectives, we analysed a corpus of fifty scientific articles written by NS of English and fifty scientific articles written by NNS of English. The variations were classified and the different occurrences counted to detect the most common ones, contrasting the different number of occurrences. The corpus we propose in this article can be used by NNS of English to avoid non-familiar terms and it also evidences the influences of mother tongue when writing a second language.

1. INTRODUCTION

Language reflects our perception of reality and the way we order and construct our reality. As Kramsch (1998: 3) explains: "Speakers identify themselves and others through their use of language; they view their language as a symbol of their social identity. […] Thus we can say that language symbolizes cultural reality." Speakers transmit their own perception of reality through language and use it to persuade, influence or manipulate. The way speakers choose different rhetorical strategies in their discourse changes the disposal of the sentence or paragraph elements, resulting in language variation.

Variations are caused because writers do not use the same language structures, terms and strategies in their communication. These differences can be clearly observed when we contrast texts of the same genre but performed by writers with different social, cultural or economic background. The internal structure of the genre within a particular professional or academic context restricts the form of the linguistic resources and the functional values they assume in discourse. This restriction is used by NNS of English to express in this language more correctly, but there are occasions when variation exists. NNS of English are used to express in their mother tongue, therefore, sometimes they copy its structures when they express in a second language.

Variation is avoided if the writer is familiar with a genre before he/she uses it. There are languages that have standardized rules to write certain genres, as for example, technical or scientific English. Anna Duszak (1997: 9) adds to this idea: "Recent insights into academic writing have shown considerable variation in text characteristics across fields, languages and cultures. […] Among the most notable differences are field-and culture-bound disparities in global organization schemata of texts." The variations of texts should not vary their interpretation; on the contrary, the main aim of language, i.e. communication, cannot be performed.

Particularly, technical language owns peculiar features associated with technical thinking, as short sentences, domain specific vocabulary and simple sand direct language structures (Carrió Pastor, 2005). Technical writing differs from other genres in being very formal and direct and so, rhetorical expressions, metaphors, colloquial expressions, etc. are avoided. As Anna Duszak (1997: 2) points out: "All this contributed to the image of a dehumanised language of science, and likewise to the image of a dehumanised writer [...] uniformity of academic writing styles was taken for granted and was accounted for in terms of objectivised research standards." Academic or technical writing possesses specific characteristics that differentiate it from other genres, as Alcaraz Varó (2000: 138-9) states. It has high semantic density or conceptual of compounded lexical units and specialized expressions that highlight objectivity, the results of the research and the hiding of authors. Eggins & Martin (2000: 336) suggest further characteristics: the use of standard syntax without abbreviations; no reference to the author of the text; the topic is considered the most relevant aspect; frequent use of incrustations; i.e. to put several subordinated sentences together and dense lexical nominal phrases with long postmodification; shortened vocabulary, with action words expressed through nouns; highly specialized vocabulary; rare adverb use and the use of terms that have specialized technical meaning.

The term *variation* is introduced in this paper as the different manifestations in the language that are not mistakes or errors (Ellis, 1997) but are the differences found in the discourse produced by writers with different linguistic and cultural antecedents, although they share the knowledge of the specialist content and academic way of expressing their thoughts. Smith and Wilson (1983: 182) mention a similar term that they call *register variation*, but they apply it to the variations produced depending on the context. This is not exactly our concept of variation, as we consider the different performances of different writers in the same kind of linguistic production and register.
These variations were analysed and recorded in a corpus that allowed us to identify the most sensible parts of the sentence to be expressed in a different way by NS and NNS of English.

2. CORPUS ANALYSIS

The need to base language studies in frequencies to obtain reliable data in linguistic research is a fact that can be observed in many previous studies. Linguistic research that has scientific rigour and is objective should be based on real data and not on intuition. A corpus allows us to investigate about language use as it provides real information about the most frequent language structures and rhetoric strategies. The deep analysis of linguistic behaviour indicates language evolution as it extracts data from average language use, not form an idyllic native speaker.
There is some intrinsic strength that regulates linguistic use, but the real problem in the instability of the linguistic system is the application of statistics and probability calculation (Malmberg, 1981: 223). Language variation can be determined considering quantitative aspects and statistic probabilities, and, in this way, the intrinsic characteristics of language can be identified. Huizhong (1985: 93) justified language corpus use in this way:

> Corpus linguistics is able to provide a better model for the description of the English language, which because of the very large amount of data involved cannot be studied directly by human observations. In language study the sampling of linguistic data is indispensable. [...] Language study must be based on sampling.

A corpus, following Huizhog, should include the highest quantity of entries in order to obtain reliable results. The bigger the corpus, the higher the possibilities to obtain consistent conclusions. There are three basic requirements to obtain a reliable corpus: first, the samples should be obtained from similar texts; second, the samples should be representative of the whole corpus, and third, the texts should be useful for our research purposes, as it is not only important to use a corpus, but it is also important to interpret the results correctly.

Corpora are used to fundament theories and ideas, providing examples that support knowledge as Hornero, Luzón and Murillo (2006) transmit in their book. The scientific and mathematical analysis of language provide credibility to linguistic research, as there are data that support research. Nevertheless, there have been some linguists that have not considered corpus analysis a useful tool for language research, as for example, Chomsky (Coulthard, 1988: 2):

> Chomsky suggested that not only was a corpus unnecessary, it was actually counterproductive. No corpus, however large, can be adequate because it will never contain examples of all possible structures and will actually contain misleading data, performance errors [...].

Nowadays, more and more researchers have accepted corpus analysis as a way to justify their research, using percentages and frequencies to analyse language use. The importance of corpora analysis and its application to applied linguistics is beyond doubt, as recent studies can confirm (Holmes, 1994; Kourilova, 1996; Ceirano and Rodriguez, 1997; Biber, Conrad and Reppen, 1998; de Monnink, 1998; Martí Guinovart, 1999; Oostdijk, 2000; Cortese, 2002; Hornero, Luzón and Murillo, 2006).

One of the recent approaches of corpus linguistics has been the lexicographic analysis of texts, lead by Sinclair with the COBUILD project in the University of Birmingham (Carter, 1998: 167; McCarthy, 2001: 125). They have designed software to classify and search lexical units in order to extract information about language use and English collocations. From then on, many other corpus projects have been developed as for example, ELRA (European Language Resources Association); ICAME (International Computer Archive of Modern and Medieval English); The Oxford Text Archive; The Cambridge International Corpus; The British National Corpus; Linguistic Resources on the Internet; IT Centers for English Linguistics Corpus; the Corpus of IULA, etc.

Another approach of corpus linguistics has been the compilation of grammar rules and sentence structures in order to identify new frames and to adapt them to their real use. It is a fact that English is influenced by language users and in environments like the Internet the massive use of English as a lingua franca produces more variation now than before the digital era. As we have stated in the previous section, variations should be identified and classified in order to incorporate them to standard language, renewing language use. Through frequency analysis we can verify the use of certain structures and it can help to determine the way language is arranged. Sinclair (1991) was conscious of this fact, and this is the reason why he wished to establish a bond between sense and structure from a lexical analysis. McCarthy (2001: 127) comments about Sinclair research:

> [Sinclair's proposal] stands as a good example of how a 'neutral' technology can throw up fundamental questions for theory, and how a practical, 'applied' problem, in this case writing a dictionary using computer evidence, can bounce back and challenge theory. We should not doubt that

> galloping technological change will bring many more such upheavals over the coming decades.

There are two ways to process a written corpus, on the one hand, the occurrences can be counted manually, and on the other hand, computer programs can be used to label categories or words and to count occurrences. The most well known computer programs are *Tagged British National Corpus* (Leech, 1997), *IULA,* (Morel, 1997), *MonoConc Pro* (Barlow, 1998) or *WordSmith* (Scott, 1998), that label corpora in order to count occurrences and frequencies previously selected. Independently of the way a corpus is processed, we use it as a way to provide data that serves as evidence for our analysis. Therefore, the advantages of using corpora to investigate are evident, but some researchers advise to pay special attention to data interpretation, as Carter (1998: 233) comments:

> Computer corpora allow access to detailed and quantifiable syntactic, semantic and pragmatic information about the behaviour of lexical items. There is little doubt that such corpora offer invaluable data for vocabulary materials development. But there are obvious dangers in using such data without carefully interpreting it as data and without careful assessment of the kinds of pedagogic criteria which might inform its use.

A well designed corpus can support our generalizations, but if the figures are interpreted erroneously, all our research is not acceptable.
In this paper, corpus analysis is going to be used to demonstrate language variation in English research articles. There are some disciplines that use well known linguistic formulae to express findings all around the world, but even in these static genres, variation exist. The objectives of this paper are to show that a corpus compiling the variations found in a standardised context as scientific articles can illustrate the parts of the sentence that are more sensible to variation and it can also make obvious the incidence of changes in a written text, evidencing the non-standardisation of language use.

3. MATERIAL AND METHOD

In order to obtain sound conclusions, 100 scientific articles from international journals were selected, 50 written by Spanish NNS of English and 50 by NS of English. Different journals, from among the most well-known in the given areas of study, were selected to generate a representative corpus. The article authors were NS from the United Kingdom or United States, and the NNS were from Universidad Politécnica de Valencia, whose articles had been revised by Spanish linguistic experts. The articles were correct, but native speakers of English have not suggested changes.
Once the research corpus was compiled, all the variations were located and counted, and percentages and frequencies were calculated in the corpus. Scott's computer program, *Wordsmith* (Oxford University Press) was used to calculate variations, counting the elements found in the sentences.
We focused on the variations found in noun phrases, verb phrases, conjunctions and epistemic modality expressions. The variations were counted and classified in tables in order to observe their occurrences and frequency. The results were analysed and we calculated χ^2 (ji-square value) in order to obtain the statistical value that is relevant if the contrasted occurrences are inferior to 0.05. Finally, the conclusions of this analysis were exposed.

4. RESULTS

The corpus gathered to analyse language variation can be observed in Table 1:

SENTENCE DATA	OCCURRENCES NNS (%)	OCCURRENCES NS (%)
Total words	184,357 (47.11%)	206,907 (52.89%)
Word list	10,590 (45.43%)	12,716 (54.57%)
Sentence number	9,017 (50.00%)	9,017 (50.00%)
Word average	20.44 (46.11%)	22.94 (53.89%)
Paragraph number	1,145 (55.51%)	916 (44.49%)
Paragraph word number	161.29 (41.58%)	225.88 (58.12%)

Table 1. Corpus gathered from English research articles.

We analysed the same sentence number in order to contrast the results obtained from NS of English and NNS of English. The first aspect we considered in the corpus analysis was the variations that obtained when contrasting noun phrases. Noun phrases are important in scientific English as they are commonly used in order to abbreviate texts, so we contrasted NS texts and NNS texts in order to observe if the most complex combinations were used in the same way by writers with different linguistic background. We also analysed the use of noun phrases followed by *of* preposition and the use of the article as complex noun phrases and articles are used in a different way in Spanish and in English. The results can be observed in Table 2:

NOUN PHRASE COMBINATIONS	OCCURRENCES NNS (%)	OCCURRENCES NS (%)	χ^2
N3	679 (53.61%)	590 (46.49%)	P = 0.14
A+ N2	906 (49.81%)	913 (50.19%)	P = 0.04
A2+ N	313 (46.58%)	359 (53.42%)	P = 0.00
N4	52 (63.41%)	30 (36.59%)	P = 0.03
A+ N3	126 (60.29%)	83 (39.71%)	P = 0.01
A2+ N2	53 (45.69%)	63 (54.31%)	P = 0.19
A3+ N	8 (44.44%)	10 (55.56%)	P = 0.53
N5	3 (60.00%)	2 (40.00%)	P = 0.70
A+ N4	12 (80.00%)	3 (20.00%)	P = 0.02
A2+ N3	7 (50.00%)	7 (50.00%)	P = 0.89
A3+ N2	1 (33.33%)	2 (66.67%)	P = 0.52
A4+ N	0 (0.00%)	1 (100.00%)	-
N6	0	0	-
Total NP	2839 (51.69%)	2653 (48.31%)	-
N+ 'OF'	4341 (46.41%)	5013 (53.59%)	-

Articles Total A AN THE	21626 (48.33%) 3965 (46.23%) 841 (48.50%) 16820(48.85%)	23113 (51.67%) 4611 (53.77%) 893 (51.50%) 17609(51.15 %)	- P = 0.00 P = 0.89 P = 0.00

Table 2. Noun Phrase variations.

The following aspect to be analysed was the variations found in the usage of verb tenses, considering also important aspects the use of modal verbs and the use of the passive voice. The occurrences found in the two groups of our corpus can be seen in Table 3:

VERB PHRASES	OCCURRENCES NNS (%)	OCCURRENCES NS (%)	χ^2
Present simple	3034 (47.71%)	3324 (52.29%)	P = 0.01
Present continuous	34 (58.62%)	24 (41.38%)	P = 0.14
Past simple	5145 (48.98%)	5359 (51.02%)	P = 0.93
Past continuous	5 (35.71%)	9 (64.29%)	P = 0.32
Present perfect	40 (42.55%)	54 (57.45%)	P = 0.21
Past perfect	1 (11.11%)	8 (88.89%)	P = 0.02
Future (will)	424 (60.65%)	275 (39.35%)	P = 0.00
Total verb tenses	8683 (48.95%)	9053 (52.83%)	-
Modal verbs	1769 (54.16%)	1497 (45.84%)	-
Passive voice	248 (43.43%)	323 (56.57%)	-

Table 3. Verb phrase variation.

In Table 4 we can observe the occurrences found in the use of modal verbs and the different usage done by native speakers of English and by non native speakers of English:

MODAL VERBS	OCCURRENCES NNS (%)	OCCURRENCES NS (%)	χ^2
CAN/ BE ABLE	877 (59.82%) 78 (76.47%)	589 (40.18%) 24 (23.53%)	P = 0.00 P = 0.00
COULD	166 (48.82%)	174 (51.18%)	P = 0.03
MAY	181 (39.69%)	275 (60.31%)	P = 0.00
MIGHT	13 (24.07%)	41 (75.93%)	P = 0.00
MUST	213 (62.64%)	127 (37.36%)	P = 0.00
NEED	90 (38.96%)	141 (61.04%)	P = 0.00
SHOULD	151 (54.51%)	126 (45.49%)	P = 0.90

Total	1769 (54.16%)	1497 (45.84%)	-

Table 4. Modal verb variation.

The analysis of conjunction usage was also considered relevant in our study as their treatment is quite different in English and in Spanish. The results obtained can be seen in Table 5:

CONJUNCTIONS	OCCURRENCES NNS (%)	OCCURRENCES NS (%)	χ^2
1. Additive	594 (37.57%)	987 (68.43%)	P = 0.00
2. Adversative	611 (46.14%)	713 (53.86%)	P = 0.00
3. Cause	408 (43.45%)	531 (56.55%)	P = 0.03
4. Time	273 (34.95%)	544 (65.05%)	P = 0.00
Total	1886 (40.52%)	2775 (59.48%)	-

Table 5. Conjunction variation.

Finally, we also considered relevant the analysis and contrast of some academic writing related to epistemic modality, i.e. words related to human judgment, as the use of abbreviations, informal words, uncertainty and certainty expressions and impersonal forms. The results obtained in this scrutiny can be seen in Table 6:

EPISTEMIC MODALITY	OCCURRENCES NNS (%)	OCCURRENCES NS (%)	χ^2
Abbreviations	32 (44.44%)	40 (55.56%)	P = 0.86
Informal words	0 (0.00%)	2 (100.00%)	P = 0.19
Uncertainty expressions	957 (48.87%)	1001 (51.13%)	P = 0.00
Certainty expressions	546 (64.38%)	302 (35.62%)	P = 0.00
Impersonal forms	959 (49.86%)	964 (50.14%)	P = 0.00

Table 6. Academic writing variation.

We did not analyse further data in our corpus because we considered that the occurrences obtained identified those parts of the sentence more sensible to language variation.

5. CONCLUSION

The main research hypothesis of this article is to identify those aspects in English language that are more susceptible to variation if we contrast texts written by NS of English and Spanish NNS of English. The corpus we obtained after analyzing the research articles of the groups of writers showed that there are certain parts of the sentence more sensible to vary when they are used by native speakers of English and by non native speakers of English. The results obtained in the use of complex noun phrases demonstrated that the longer the complex noun phrases, the more variation we could find. Native speakers of English used less complex noun phrases formed by four or five

elements than Spanish non native speakers of English. These results were quite surprising as complex structures are in general more difficult to use by NNS, but it could be caused by an overuse of recommended structures in technical English. Also, we have to notice that the use of noun phrases followed by *of* was more common among NS of English, so the use of complex noun phrases in technical English is not so common in NS than in NNS of English. The NS also used more articles than NNS, due to the different concept of the article in Spanish and in English. Consequently, writers should pay special attention to article usage and not to overuse complex noun phrases.

The verb tenses examined demonstrated that most of the tenses were used in the same manner, but we should pay special attention to the use of the future form *will*. This form was more used by NNS of English than by NS of English, as a result of the different conception of the *will future* in English and in Spanish. In Spanish, the future form expresses certainty; meanwhile in English it expresses uncertainty. We consider that Spanish writers apply mother tongue patterns in the use of this tense and this is the cause of its overuse. We can also determine that the different use of the passive voice is also due to the influence of Spanish. The passive voice in Spanish is not used as an impersonal form, but in English it is usually used in scientific English to express research findings.

We considered modal verb frequencies in more detail as they indicate modality in English and the intention of the writer. If we observe the results shown in Table 4, we can notice that Spanish non native speakers used *can, be able* and *must* in more sentences than English speakers. On the contrary, the latter used more *may* and *might* than Spanish writers. These results confirmed that there are several parts of the sentence more sensible to variation, and the ones related to certainty or uncertainty are used in a different way by both groups of writers. Spanish language expresses ideas or findings assertively, whereas English language prefers to use other language strategies.

The results obtained in the use of conjunctions confirmed that English writers use more conjunctions than Spanish writers, but this could be considered a lack of language proficiency to join ideas.

Finally, the results obtained in certain parts of the sentences related to epistemic modality revealed native speakers of English and non native speakers of English used abbreviations, informal words, uncertainty expressions and impersonal forms with the same frequency, but we can observe in Table 6 that certainty expressions are more used by Spanish NNS of English. This confirms the data obtained in other sentence parts as the use of *will* or certain modal verbs. Spanish writers of English express certainty in their research articles as a result of the influence of their mother tongue. This variation can be observed in different parts of the sentences analysed.

This finding evidences the need to incorporate language variations to the English language. These variations should be identified in a multimodal international corpus that showed the changes in language production incorporated by second language speakers. The detection of these variations would allow linguists to incorporate mother tongue influences in English as an internationally recognized language.

6. REFERENCES

Alcaraz Varó, E. (2000). *El inglés profesional y académico.* Madrid: Alianza.

Barlow, M. (1998). "MonoConc concordance programs for text analysis". Http://www.ruf.rice.edu/barlow/mono.html (03-06-07)

Biber, D.; Conrad, S. y Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Carrió Pastor, M. L. (2005). *Contrastive analysis of scientific-technical discourse: common writing errors and variations in the use of English as a non-native language.* Ann Arbor: UMI.

Carter, R. (1998). *Linguistic Choice across Genres: Variation in Spoken and Written English*. Amsterdam: John Benjamins.

Ceirano, V. & Rodriguez P. G. (1997). "Análisis del discurso asistido por computadora. Nuestra experiencia con el NUD IST". Http://www.geocities.com/Athens/ Forum/5917/analista.html (06-04-08)

Cortese, G. (2002). "My 'Doxy' Is not your 'Doxy': Doing corpus linguistics as collaborative design" in G. Cortese & P. Riley [eds.] *Domain-specific English.* Bern: Peter Lang: 367- 414.

Coulthard, M. (1988). *An Introduction to Discourse Analysis.* London: Longman.

Duszak, A. (1997). "Cross-cultural academic communication" in A. Duszak [ed.] *Culture and styles of academic writing*. New York: Mouton de Gruyter: 11- 40.

Eggins , S, & Martin, J. R. (2000). "Géneros y registros del discurso" in T. A. Van Dijk [ed.] *El discurso como estructura y proceso*. Barcelona: Gedisa Editoral: 335- 371.

Ellis, R. (1997). *Second Language Acquisition.* Oxford: Oxford University Press.

Holmes, J. (1994). "Inferring language change from computer corpora: Some methodological problems". *ICAME Journal. Computers in English Linguistics*, 18: 27- 40.

Hornero, A. M.; Luzón, M. J. and Murillo, S. [Eds.] (2006). *Corpus Linguistics. Applications for the Study of English.* Bern: Peter Lang.

Huizhong, Y. (1985). "The use of computers in English teaching and research in China" in R. Quirk & H. G. Widdowson [eds.] *English in the World. Teaching and Learning the Language and Literatures.* Cambridge: Cambridge University Press: 86- 104.

Kourilova, M. (1996). "Interactive functions of language in peer reviews of medical papers written by non-native users of English". *UNESCO-ALSED-LSP Newsletter,* 19- 1: 4- 21.

Kramsch, C. (1998). *Language and Culture.* Oxford: Oxford University Press.

Leech, G. (1997). "A brief user's guide to the grammatical tagging of the British National Corpus". Http://www.hcu.ox.ac.uk/BCN/what/gramtag.html (03-06-02)

Malmberg, B. (1981). *Los nuevos caminos de la lingüística.* Madrid: Siglo XXI editores.

Martí Guinovart, M. A. (1999). "Panorama de la lingüística computacional en Europa". *Revista Española de Lingüística Aplicada. Volumen Monográfico: Panorama de la Investigación en Lingüística Aplicada*, 11- 24.

McCarthy, M. (2001). *Issues in Applied Linguistics.* Cambridge: Cambridge University Press.

Mönnink, I. de (1997). "Using corpus and experimental data: a multimethod approach". Http://iris1.let.kun.nl/literature/demonnink.1997.2.html (13-03-98, 12:26)

Morel, J. (1997). *El Corpus de l'IULA: etiquetaris.* Barcelona: Universitat Pompeu Fabra.

Oostdijk, N. (2000). "Corpus-based English linguistics at a cross-roads". *English Studies*, 81- 2: 127- 141.

Scott, M. (1998). *WordSmith Tools 3.0.* Oxford: Oxford University Press. http://www.liv.ac.uk/ms2928/wordsmit.htm.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.
Smith, N. & Wilson, D. (1983). *La lingüística moderna.* Barcelona: Editorial Anagrama.
Stubbs, M. (1995). *Texts and Corpus Analysis: Computer- Assisted Studies of Language and Culture*. Cambridge: Blackwell.
Widdowson, H. G. (2000). "On the limitations of linguistics applied". *Applied Linguistics*, 21- 1: 3- 25.