A multivariate analysis using English and Spanish corpora

Marta Sáiz*
Department of Language Engineering, UMIST
PO Box 88, Manchester M60 1QD, U.K.
Tlf: +44.161.200.3078
Fax: +44.161.200.3099

Abstract

In this paper we describe a study to carry out a multivariate cross-linguistic analysis of linguistic patterns across specialized parallel corpora in English and Spanish using factor analysis. First, we extracted the factors by reducing the original set of variables, formed by linguistic features, to a smaller number of factors and interpreted each factor taking into consideration the functions shared by the underlying co-occurring patterns. Then we carried out cross-linguistic comparisons of English and Spanish by comparing the identified dimensions with respect to their underlying functions, associated linguistic characteristics and the multidimensional characterization of registers included in this study.

1 *Introduction*

Quantitative approaches to corpus analysis are important to the development of tools for applications in Natural Language Processing. They allow us to examine reliably the frequency of specific linguistic patterns of use in texts and also to classify them and even to construct statistical models with a view to expalining what we observe in the data.

Factor analysis is a quantitative multivariate technique which allows the examination of complex relationships among a set of interrelated variables. Multivariate techniques are widely used in many fields of research. They allow manipulation of many variables in a single analysis and are available in a number of computer packages. The one we have used is SPSS (Statistical Package for the Social Sciences), running the FACTOR procedure using principal component analysis. Our use of factor analysis is exploratory and involves detecting patterns of variables to discover new concepts and a possible reduction of the data.

The multivariate approach to linguistic variation was first used by Biber (1985) and more completely developed by Biber (1988) to describe the textual relations among spoken and written genres. In his 1988 study, he examined 67 different linguistic features in 23 different text types. The linguistic features were counted in the texts (comprising about one million words in all), and the quantitative results were submitted to factor analysis. The 67 features were reduced to seven factors and interpreted in terms of the communicative functions shared by the co-occurring linguistic features.

Biber (1988, 1995) has proven that cross-linguistic register similarities or differences cannot be analyzed in terms of a single parameter or dimension, of simple dichotomous distinctions or by analyzing registers identifying the salient co-occurrence patterns on an intuitive basis. Rather, the use of quantitative techniques in multivariate studies empirically identifies sets of linguistic features that co-occur in registers, and the different shared functions underlying these co-occurring patterns are interpreted by considering previous research on text analysis and by performing qualitative analysis of particular texts.

Our study follows a similar general methodology — we carry out a *cross-linguistic* analysis of linguistic patterns across specialized texts in English and Spanish using parallel texts. The objective of this study is twofold: one is to make progress in cross-linguistic comparisons by better describing particular registers in different languages. If the full range of linguistic variability in registers is described, systems can be developed which deal with that range of variation. Carrying out this kind of analysis is useful for NLP, e.g. it is highly relevant to analyze bilingual corpora to develop MT systems. Our other objective with this study is to analyze the identified dimensions in English and Spanish with respect to their underlying functions, their associated linguistic characteristics and the relations among the registers included in the present study with respect to the identified dimensions, in order to investigate how language operates in translations. If the dimensions found are similar across English and Spanish with respect to these just mentioned aspects, this should then be a point to consider when developing computational systems for MT. For this reason, 4 of the 5 corpora included in this study are parallel translations in English and Spanish, the other is parallel in the sense that it consists of law texts in both languages but not translations (i.e. it is a comparable corpus).

* Research sponsored by the Departamento de Educación, Universidades e Investigación of the Basque Government.

2 *Methodology*

The following steps were carried out prior to the submission of quantitative data to factor analysis (FA):

- Collection of texts: we selected parallel texts in English and Spanish from 5 different corpora[1] from the European Corpus Initiative CD-ROM, comprising 230 texts for English, 223 for Spanish and about 225,000 words in all for both.
- Submission of texts to a POS tagger to identify linguistic features in texts. For English, Brill's tagger was used (Brill, 1992); for Spanish, we trained Brill's tagger to deal with this language. Our POS tagger has been trained following the same technique originally developed by Brill, called transformation-based-error-driven learning. The tagger automatically acquires rules from tagged text increasingly improving its performance. In all, 37 morphosyntactic features were included for English and 38 for Spanish. In order not to rely uniquely on the morpho-syntactic features which the POS tagger recognizes, we also considered for our study measures typically used in stylometric studies (average word length of sentences, average length of words, long words, average number of sentences per text and type/token ratio) in the belief that including these would help to discover genuine differences in the texts belonging to different registers we have analyzed. These stylistic measures were extracted using WordSmith tools.
- Manual revision of the tagged English and Spanish output to correct any errors made by the tagger.
- Counting of the linguistic features in texts and normalization of frequency counts to a text length of 1000 words. As text length varies for all texts, for a fair comparison of frequency counts of linguistic features across texts, normalization of frequency counts was necessary.
- Submission of normalized counts to FA so that the large number of original variables are reduced to a smaller set of derived variables: the factors, typically formed by sets of co-occurring linguistic features in texts.

There are several steps involved in FA. The starting point is the examination of the correlation matrix of the variables and samples; then the factors necessary to represent the data are extracted. The method for factor extraction used was principal components analysis which transforms original variables into a new set of uncorrelated variables with decreasing amounts of variation. The new variables are principal components rather than factors although for convenience we will call them factors. Rotation of the factors to a simpler structure then takes place to make them more interpretable. Lastly, the factors are interpreted functionally in terms of the communicative functions shared by the co-occurring features, determined by looking at previous research on the literature (Chafe, 1982; Biber, 1988 and Biber, 1995) and by carrying out a detailed qualitative analysis of the linguistic features in texts.

3 *Factor structure*

In this multidimensional study, 5 factors were extracted for both Spanish and English as this was considered the optimal among several solutions considered[2]. The factors extracted identify a group of linguistic features which frequently co-occur in texts. It is claimed that the linguistic features in these groups co-occur because they serve a common underlying communicative function (Biber, 1988). Thus, to identify the textual dimensions underlying each factor, the factorial structures for both languages were interpreted taking into account the communicative functions shared by the linguistic features which loaded on each factor. Each factor is given a label optimally describing these communicative functions. Table 1 summarizes the identified dimensions, the interpretative labels which explain the functions associated with each dimension and the linguistic features grouped on each factor (in brackets). Linguistic features can be positive or negative, however, this does not indicate that positively-valued linguistic features are more important. Rather, it shows groups of features which co-occur in complementary distribution. For instance, in EN-D1, among the positive linguistic features included are: non third person of verbs, gerunds, type/token ratio, adverbs (comparative and superlative), infinitive form of verbs, 3rd person singular present verbs, the feature 'to', etc., while there is only one negative-valued linguistic feature: cardinal number. Referring to dimension 1, this means that when the features with positive value occur it will be likely that the features with negative values will not occur and vice versa. Discussions in the research literature and also our own functional analysis of individual features in the texts considered for this study provided the

[1] The announcement text of the Esprit R&D programme, the Xerox ScanWorX User's Guide, International Labour Organization Reports on the Committee on Freedom of Association, the International Telecommunications Union C-CITT Blue Book and the Copenhagen Business School Civil Law Corpus.

[2] This decision was made considering several guidelines generally used when determining the number of factors to be extracted, such as observing a plot with the total variance associated with each factor. Such a plot is called *scree plot* and typically shows the point at which additional factors contribute little to the overall analysis.

tivation for interpreting these features in terms the given dimension labels.

the following sections, we carry out cross-guistic comparisons of English and Spanish by mparing the dimensions across both languages ith respect to their underlying functions, their guistic characterisitics co-occurring in each di-ension and their relations with respect to the re-isters included in our study.

3.1 Cross-linguistic comparisons of dimensions with respect to their underlying functions

As can be seen from Table 1 the majority of dimensions have close similarities across English and Spanish with respect to their underlying functions and there are communicative functions shared across the two languages. Among the dimensions with functional correspondences is dimension 1 in English and dimension 5 in Spanish[3], both of which are interpreted as reflecting "Descriptive Production vs Non-Descriptive Production". Another communicative function shared across the two languages is narration, represented in SP-D2 and in EN-D2 and in the positive pole of EN-D4. This is a more complex correspondence, since there is only one dimension in Spanish which reflects narration and which corresponds to two dimensions in English. There are other dimensions with functional correspondences across English and Spanish: one is located in the negative pole of EN-D4 and the positive pole of SP-D2, labelled as "Informational Persuasive Production"; another in the negative pole of EN-D5 and SP-D4, interpreted as "Procedural/Direct Production" and "Procedural vs Non-Procedural Production" respectively; and another in the positive pole of English EN-D5 and SP-D4, interpreted as "Elaborate Discourse" and "Elaborate Discourse vs Unelaborate discourse" respectively.

As can also be seen from Table 1, some communicative dimensions are specific to English while others are specific to Spanish: this is the case of EN-D3, interpreted as "Specialized Information vs Non-Specialized Information" and SP-D1, interpreted as "Non-Specific Informational Reference vs Specific Informational Reference". There are two explanations for the occurrence of these communicative functions specific to a language. One is that the linguistic features which, by way of example, co-occur in EN-D3, and which reflect the production of Specialized Information, are appropiate for English but not for Spanish, for which case a different set of linguistic features, or a more refined one, might be more appropiate for the identification of such a communicative function. However, this is an issue which can only be further investigated by conducting a multidimensional analysis on a larger scale, for instance, considering a more refined set of linguistic features and possibly also more registers. This is not our task here and therefore until this is further examined we cannot check the veracity of such a claim. Another more plausible explanation can be offered by arguing that those communicative functions specific to a language reflect the communicative priorities of each language or culture (Biber, 1995:264).

[3]For better readability purposes we use henceforth EN-D1 to refer to English dimension 1, SP-D5 to refer to Spanish dimension 5, etc.

3.2 Cross-linguistic comparisons of dimensions with respect to their linguistic characteristics

We can also compare dimensions according to their defining linguistic characteristics. As discussed below, dimensions with corresponding underlying functions have similar associated linguistic features, such as the dimensions with the characteristics of Narrative Production, Procedural Production, Elaboration, and Informational Persuasive Production. With respect to Narrative Production, EN-D2 and EN-D4, and SP-D2 are related in their occurring linguistic characteristics: we note that participles, past tense verbs, type/token ratio, first person pronouns and proper singular nouns all occur in both languages. Those dimensions which reflect Procedural Information, SP-D4 and EN-D5, indicate that there is an addressor required in both languages, who gives the procedures, using the infinitive and imperative form (merged in English under the label "VB") and who addresses directly the addressee, by using second person pronouns, which represent the potential person who is to carry out the instructions or procedures; considering Elaboration, both languages are characterized by long sentences, use of prepositions and prepositional phrases, participles and relatives (indirect relative pronouns in English and possessive relative pronouns in Spanish); as for Informational Persuasive Production, EN-D4 and SP-D2 are both represented by singular nouns and modal verbs, which are the strongest features of these two dimensions. Also, to a lesser extent, those dimensions reflecting Descriptive Information (EN-D1 and SP-D5) tend to be related in their occurring linguistic features: we note both languages make use of adjectives, adverbs, coordinative conjunctions and a high type/token ratio, apart from other linguistic features used specifically by each language, to express descriptive information.

3.3 Cross-linguistic comparisons of registers

Registers can be compared with respect to each dimension by computing factor scores (Biber, 1988): for each text the frequency of each of the features having salient loadings on a factor is summed, then the factor score means of each register are computed to obtain mean dimension scores for each register on each dimension. These mean dimension scores can then be compared to define the relations among

Table 1: *Summary of English and Spanish functions*

Dimension	*English function (Linguistic features)*	*Spanish function (Linguistic features)*
Dim. 1	**Descriptive Production** *versus* (VBP/VBG/TTR/RB/VB/VBZ/TO/RBR JJS/RBS/PHR/SYM/PPI/CC/SECP) **Non-Descriptive Production** (CD)	**Non-specific Informational Reference** *versus* (RB/WP/VBZ/CC/THIRDP/PP/FW/PPI) **Specific Informational Reference** (NNPS/NNS/NNP/NUMS/CD)
Dim. 2	**Narrative Production** *versus* (DT/VBN/IN/VBD/WP/PDT/WPS/TTR) **Non-Narrative Production**	**Informational Persuasive Production** *versus* (NN/VBM/NUMS) **Narrative Production** (VBD/VBC/TTR/FIRSTP/AWL/NNPS) THIRDP/NNP/ALW/VBN)
Dim. 3	**Specialized Information** *versus* (AWL/NNS/LONGW/JJ/LS/CC/NNPS/IN) **Non-Specialized Information** (RB/VB/VBZ/POS/EX)	**Elaborate Discourse** *versus* (IN/DT/LONGW/VBN/ALW/WPS/NNS/JJ) **Unelaborate Discourse** (CD/NUMS)
Dim. 4	**Interpersonal Narrative** *versus* (NNP/POS/FIRSTP/VBD) **Informational Persuasive Production** (NN/MD/VB/JJ)	**Procedural Production** *versus* (VB/VBI/SECP/SYM/VBG/WRB/RBR/JJS) **Non-Procedural Production** (LS)
Dim. 5	**Elaborate Discourse** *versus* (THIRDP/ALW/EX/FW/WRB/TTR/VBN IN/CC) **Procedural Production** (NUMS/SECP/VBP/VB)	**Descriptive Production** *versus* (JJ/JJR/VBF/RBS/CC/NNS/TTR) **Non-Descriptive Production** (INDT/NNP/VBD)

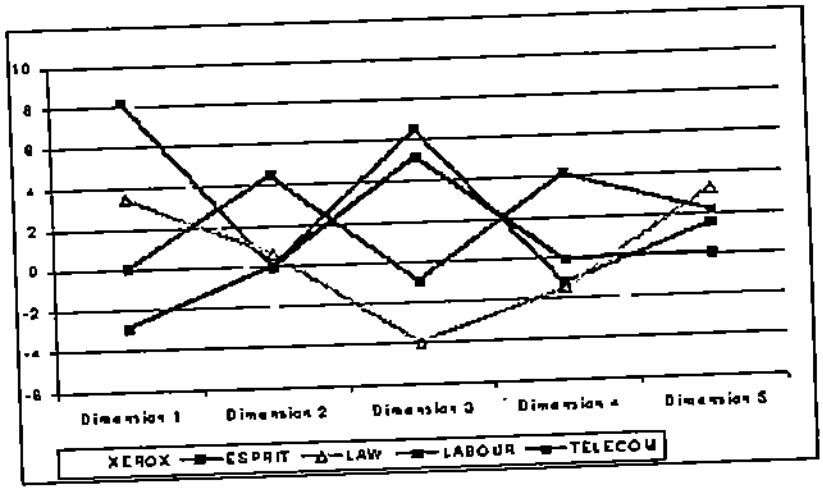

Figure 1: Multidimensional profile for English texts

registers. In Figure 1, the mean dimension score for each register for each dimension for English is plotted. For Spanish, analogous information is shown in Figure 2. We can see that registers have different characterizations with respect to each dimension and that it is only by considering the 5 dimensions that we can detect comparisons or differences among the registers and that we are able to describe these in a motivated manner.

We have further found that dimensions which are similar from a functional and linguistic perspective also tend to be similar with respect to their relations with the registers. There are major register similarities in English and Spanish regarding their overall characterization with respect to the communicative functions defined by the dimensions. The most straightforward of these involve the narrative and procedural functions. The same parallel registers, LABOUR and XEROX, are highly marked in both languages for the narrative and procedural functions respectively. ESPRIT is highly descriptive in the two languages. There are also more complex correspondences. Elaboration is markedly present in English for LAW, followed by LABOUR and ESPRIT while in Spanish elaboration is represented by LABOUR and ESPRIT. However, the LAW corpus, even though it refers to law texts in English and Spanish, is the only non-parallel corpus included in this study. The fact that LAW is elaborate in

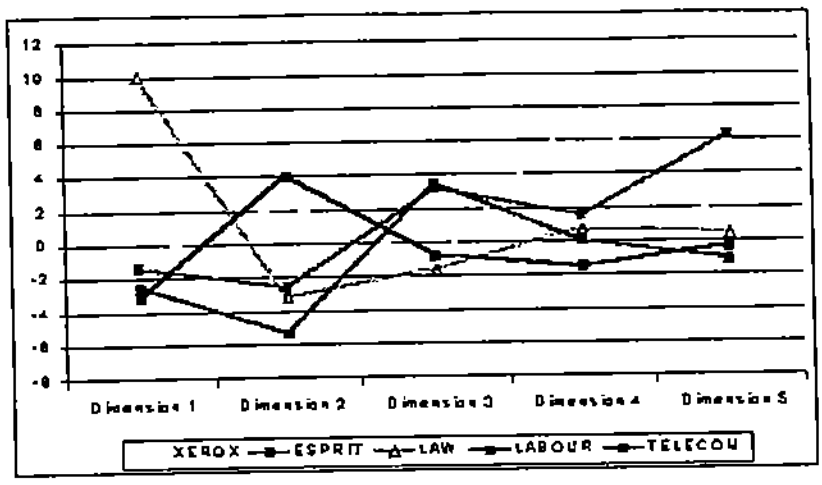

Figure 2: Multidimensional profile for Spanish texts

English and non-elaborate in Spanish provides strong evidence to suggest that parallel texts tend to be more similar cross-linguistically than non-parallel texts and that therefore language in translations tends to keep similar structures.
There are still some differences which show that parallel registers differ in English and Spanish, in terms of our not having found identical multidimensional characterization of registers across the two languages; for example, the informational persuasive function is stronger in TELECOM and XEROX in Spanish and in LAW and ESPRIT in English, a difference caused by the overall allocation of linguistic features to this communicative function.

4 *Conclusions*

The cross-linguistic comparisons of dimensions undertaken above with respect to their underlying functions, associated linguistic characteristics and the multidimensional characterization of registers, have provided evidence for the following findings:

- A multidimensional analysis is required, focussing on the five communicative dimensions identified in the present study, in order to capture the complexity of the linguistic relations in the registers considered in English and Spanish.
- The cross-linguistic similarities found are stronger than the cross-linguistic differences.
- The majority of the dimensions have corresponding (or analogous) underlying functions: we note both English and Spanish have dimensions which reflect description, narration, procedural production, elaboration and informational persuasive production.
- Dimensions with corresponding underlying functions also tend to have similar linguistic features associated with them. One significant example is that of the procedural function: in both languages there is an addressor who tells how to carry out the instructions or procedures using the imperative and the infinitive form and there is an addressee, who carries the procedures out and who is referred to by the use of second person pronouns.
- The same registers in English and Spanish tend to be clearly marked for similar communicative functions, for example ESPRIT, XEROX and LABOUR are highly descriptive, procedural and narrative respectively in the two languages.
- Parallel texts tend to be more similar cross-linguistically than non-parallel texts. LAW is the only non-parallel register included in this study. We found that, unlike the other registers, which tend to be characterized multidimensionally in more analogous ways, LAW presents more differences with respect to its multidimensional characterization in relation to the communicative functions defined by the dimensions. Especially relevant is the fact that LAW is clearly elaborate in English and non-elaborate in Spanish. This provides strong evidence for suggesting that language in translations tends to keep similar structures.

5 *References*

D. Biber. 1985. Investigating macroscopic textual variation through multi-feature/ multidimensional analyses. *Linguistics*, 23:337-60.
D. Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.
D. Biber. 1995. *Dimensions of Register Variation*. Cambridge University Press.

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*. ACL.
W. Chafe. 1982. Integration and Involvement in speaking, writing, and oral literature. In Tannen, editor. *Spoken and written language: exploring orality and literacy*. pp. 35-54. Ablex.

Appendix A ENGLISH TAGSET

1.	CC	Coordinating Conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existencial 'there'
5.	FW	Foreign word
6.	FIRSTP	1st person pronouns
7.	IN	Preposition or subordinating conjunction
8.	JJ	Adjective
9.	JJR	Adjective, comparative
10.	JJS	Adjective, superlative
11.	LS	List marker item
12.	MD	Modal
13.	NN	Noun, singular or mass
14.	NNS	Noun, plural
15.	NNP	Proper noun, singular
16.	NNPS	Proper noun, plural
17.	PDT	Predeterminer
18.	PHR	Expression
19.	POS	Possessive ending
20.	PPI	Indefinite pronoun
21.	RB	Adverb
22.	RBR	Adverb, comparative
23.	RBS	Adverb, superlative
24.	SECP	2nd person pronouns
25.	SYM	Symbol
26.	THIRDP	Third person pronouns
27.	TO	'to'
28.	VB	Verb, base form
29.	VBD	Verb, past tense
30.	VBG	Verb, gerund or past participle
31.	VBN	Verb, past participle
32.	VBP	Verb, non-3rd person singular present
33.	VBZ	Verb, 3rd person singular present
34.	WDT	Wh-determiner
35.	WP	Wh-pronoun
36.	WPS	Possessive wh-pronoun
37.	WRB	Wh-adverb
38.	TTR	Type/token ratio
39.	LONGW	Long word
40.	NUMS	Number of sentences
41.	AWL	Average word length
42.	ALW	Average length of words (per sentence)

Appendix B SPANISH TAGSET

1.	CC	Coordinating Conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	FW	Foreign word
5.	FIRSTP	1st person pronouns
6.	IN	Preposition or subordinating conjunction
7.	INDT	Contraction
8.	JJ	Adjective
9.	JJR	Adjective, comparative
10.	JJS	Adjective, superlative
11.	LS	List marker item
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PHR	Expression
17.	PP	Demostrative pronoun
18.	PPI	Indefinite pronoun
19.	RB	Adverb
20.	RBR	Adverb, comparative
21.	RBS	Adverb, superlative
22.	SECP	second person pronouns
23.	SYM	Symbol
24.	THIRDP	Third person pronouns
25.	VB	Verb, infinitive
26.	VBC	Verb, conditional
27.	VBD	Verb, past tense (preterit and imperfect)
28.	VBF	Verb, future tense
29.	VBG	Verb, gerund or past participle
30.	VBI	Verb, imperative
31.	VBM	Verb, modal
32.	VBN	Verb, past participle
33.	VBS	Verb, subjunctive
34.	VBZ	Verb, present tense
35.	WDT	Wh-determiner
36.	WP	Wh-pronoun
37.	WPS	Possessive wh-pronoun
38.	WRB	Wh-adverb
39.	TTR	Type/token ratio
40.	LONGW	Long word
41.	NUMS	Number of sentences
42.	AWL	Average word length
43.	ALW	Average length of words (per sentence)