A MULTIDIMENSIONAL CORPUS-BASED ANALYSIS OF ENGLISH SPOKEN AND WRITTEN-TO-BE-SPOKEN DISCOURSE

Javier Gómez Guinovart
Javier Pérez Guerra
Facultade de Filoloxia e Traducción
Universidade de Vigo

ABSTRACT

This paper constitutes an attempt to gain some insight into the potential stylistic variables which exert some influence on the nature of spoken and written-to-be-spoken English texts. By using evidence from c. 3 million words from the British National Corpus, we apply Biber's multidimensional model to both transcriptions of actual English productions and to written material intended to be uttered by a speaker. The starting point in this investigation has been the assumption that textual samples can be characterised in functional terms by placing them on scales determined by sets of co-occurring linguistic features. Such scales are subsequently associated with functional interpretations which are decided upon the examination of the significant linguistic features.

KEY WORDS: *corpus linguistics, corpus-based analysis, spoken English and discourse analysis*

RESUMEN

Este artículo constituye un intento de adentrarse en el estudio de las posibles variables estilísticas que influyen de algún modo en lo naturaleza de los textos ingleses hablados y escritos para ser hablados. Gracias a los datos obtenidos del examen de casi 3 millones de palabras del British National Corpus, aplicamos el modelo multidimensional de Biber tanto o transcripciones de producciones lingüísticas reales en inglés como a materiales escritos para ser leídos en voz alta. El punto de partida of este estudio ha sido la asunción de que es posible caracterizar los textos en términos funcionales mediante su localización en escalas determinadas por conjuntos de rasgos lingüísticos que tienden u coaparecer en los textos. Estas escalas se asocian posteriormente con interpretaciones funcionales resultantes del examen de los rasgos lingüísticos significativos.

PALABRAS CLASE: *lingüística del corpus, análisis basado en corpus, inglés oral y análisis discursivo*

I. INTRODUCTION

The application of computational techniques to the study of the stylistic characterisation of a given text type or genre has been one of the major goals of so-called Corpus Linguistics in the last years. In this paper we embark on the stylistic examination of an electronic collection of present-day English spoken and written texts. The theoretical framework on which our analysis draws is Biber's (1988) multifeature multidimensional model.

What follows is based on the assumption that stylistic issues are governed by theory-independent concrete linguistic facts. In a nutshell. the basic idea is that the results of the quantification of the occurrence or productivity of such features in a text or group of texts can lead to the stylistic characterisation of such material. That stated. in this paper we examine a broad selection of texts with two characteristics in common: the medium and the producer. More specifically. two kinds of texts will be brought into pla';: written texts produced with a clear intention. namely. individual oral production. and spoken/oral texts produced by only one speaker. Broadly speaking. such textual categories could be grouped into the class of monologue. the only crucial difference between them being the variable ± spontaneit';. The primar? hypothesis which we attempt to either corroborate or refute is 'there exists actual grounds on which the distinction between written-to-be-spoken and spoken linguistic productions can be based'. Once such hypothesis has been given enough credence. we shall concentrate on the investigation of the Iactors which have paved the way for the characterisation of written-to-be-spoken and spoken material as two distinct categories.

This paper is organised into 4 sections. In section 2 we describe the corpus of texts. Section 3 deals with the methodological assumptions which constitute the backbone of the multidimensional model. The discussion of the framework is accompanied by the actual data obtained through the automatic analysis of the textual material. In section 4 we analyse the consequences which the application of a multidimensional model has for the functional characterisation of the texts. This section describes the dimensions resulting from the statistical process and shows the relative weight of each of the text types considered in this study in each ofthe dimensions. Finally. in section 5 we outline our conclusions and final remarks.

lI. THE CORPUS

The textual material under discussion has been extracted from the Rritish National Corpus (BNC).[1] a major electronic collection of 100 million words of actual present-day Rritish

[1] The BNC material has beeii used in this study uiider a licence issued by the BNC Consortium to the research group funded by the Spanisli Ministry of Education through its Dirección General de Enseñanza Superior (DGES). grant number PB97-0507 (licensee Professor 'Teresa Fanego). This grant is hereby gratefully achnowledged.

English dating 1960 onwards. The corpus. which is the result of a joint project involving Oxford University. Longman. Chambers Harrap. Lancaster University and the British libran. contains not only raw text but also SGML annotation.[2] Such tagging. in particular the elements 'medium' and 'interaction'. has been extremely useful since it has provided the basis for a reliable classification of the texts into spoken and written. More specifically. our subcorpus coniprises even sample tagged as 'wrimed5' with respect to the SGML-element 'medium' or 'spologl' with respect to the element 'interaction'. Put differently. we have extracted 2.890.754 words of. on the one hand. written-to-be-spoken material ('wrimed5') or 'written' henceforth. and. on the other. spoken samples produced by one speaker (monologue 'spologl') or 'spoken' onwards. Table 1 sketches the basic statistical and typological details of the corpus:

Table 1: The corpus

Major categories	taxonomy	ord totals
written-to-be-spoken	prayers	
	sermons	
	speeches	
	TV news	
	applied science	
		1.370.870
spoken	resporises	
	lectures	
	coiirses. presentations	
	talks	
	news	
	sermons	
	speeches	
	interviews	
	meetings	
	debates	
	court/parliament cases	1.519.884
Total		2.890.754

The textual taxonomies outlined in Table 1 are based on the cataloguing details given in Burnard (1995). Since the consideration of subtle differences among textual categories lies outside the scope of this paper. on a few occasions we have opted for grouping samples which belong to different categories according to the oficial classification of the corpus in order to lower the number of subcategories.

The word totals pictured in Table 1 warrant the principle of corpus representativeness since the whole length of our corpus surpasses coiisiderably that of the multidimensional studies of which we are aware. To cite a few examples. 960.000 words are used in Biber (1988): approximately 153.000 words in Besnier (1988): circa 700.000 words in Biber & Fiiiegan (1989): Atkinson's (1992) examination of medical research writing was done on a 186.553-word corpus: less than 300.000 words are used in Biber & Hared (1994): Kim &

[2] The onnotation schema adopted by the compilers of the *British National Corpus* has been CDIF (Corpus Document Interchange Format), which heavily draws on TEI (see Aston & Bumard 1997:25).

Biber's (1994) analysis of register variation in Korean is based on a corpus of 135.800 words: finally. González-Álvarez & Pérez-Guerra (1997) and (1998) explore 47.736 and c. 100.000 words. respectively. As Biber (1995a:364) himself has claimed. "the dimensions of variation (...) can be replicated in much small corpora"; in other words. the multidimensional model is not based on the length of the corpora under examination but on the range ofvariation which the textual samples exhibit.

III. THE METHODOLOGY

The starting point in Biber's n~ultidimensionalmodel is the selection of linguistic features which can be searched for and quantified in the actual textual material. By means of statistical techniques which we describe below. the normalised ratios of each of the features identified in the corpus are. íirst. filtered and. second. grouped into clusters. according to positive or negative feature occurrence. Finally. the interpretation of the resulting clusters or groups of features. on the one hand. and the scaling of the samples along the clusters. on the other. lead to. in Biber's (1994:32) words. the "overall situational and linguistic characterisation for each register".[3]

In this section we describe the multidimensional model step by step by means of our own investigation on the multidimensional characterisation of our corpus of spoken and written-to-be-spoken texts. Section 3.1. deals with the selection of the linguistic features. In section 3.2. we focus on the computational and statistical process involved in the model.

III.1. Linguistic features

Biber's initial framework involved the selection of solely linguistic features which could be observed superficially in the texts under examination.[4] In Biber (1988). for example. he identified 67 significant features corresponding to only grammatical categories which were easily discernible in a given text. In fact. Biber designed his own tagger in an attempt to ease the process of feature counting. which was based on superficial distributional facts (see. in this

[3] In this paper we shall use 'register'. 'genre' and 'text type' indistinctly. For the definition of these terms. see Biber (1994:51-53).

[4] The model is considerably enriched in Biber (1994). In this paper. Biber adds so-called (not strictly linguistic) 'situational' features such as purpose of tlie texts, social relations. production constraints. etc

connection. Biber 1988: Appendix II). This is the state of affairs to which we shall stick in this paper.

For the purposes of this pilot study. we have analysed 48 linguistic features in each text. grouped into 12 categories: (A) lexical specificity. (B) sentence length. (C) readability grades. (D) lexical classes. (E) place and time adverbials. (F) coordination and subordination. (ti) verbal forms. (H) modals. (1) some specific spntactic constructions. (J) pronouns. (K) prepositions. adjectives and adverbs. and (L) nominal categories. In order to automate the processing of these features out of the SGML tagging of the BNC. we have developed a set of computer applications in the AWK programming language.[5] To illustrate them. in Appendix 1 we show the AWK program designed to calculate the number of lexical types in a BNC text before the computation of the type-token ratio. In what follows. we discuss the linguistic features used in our work.

(A) LEXICAL SPECIFICITY

(1) type-token ratio [*ttr*]

Type-token ratio (TTR) is an index of lexical diversity (also called 'lexical density'). The TTR percentage is calculated by dividing the number of lexical types (or 'different' words in a text) by the number of lexical tokens (or text length in words). and then multiplying the result by 100. which gives the mean of different words per one hundred words of text. The reliability of TTR as a quantitative indicator of style is constrained because of its dependence on text size – while text length is theoretically unlimited. the number of different words in use in a language is finite (Holmes 1994:92). Restricting the number of words to be analysed to a limited fixed text size (disregarding the total length of the text) may improve TTR relevance (Biber 1988:238-39).

(2) *hapax legomena* [*hleg*]

In terms of vocabulary fiequency distribution. once-occuriing words or *hapax legomena* constitute. broadly speaking. the most frequent words in a text. As quantitative indicators of stgle. theg are related to vocabulary richness and precision. and have been widely used in stylometric studies on authorship attribution (Morton 1986. Holmes 1994:97-98).

(3) *hapax dislegomena* [*hdis*]

---

[5] See Bambrook (1996: Appendices 2-3) for an introduction io tlir possible application of AWK in thr humanities.

Another relevant index of vocabulary richness is achieved by counting the nunibei of *hapax dislegomena* or words which occur only twice in a text. Note that in our coniputation of lexical distribution (TTR. *hapax legomena* and *dislegomena*) we have not regarded homographs as different words because their frequency is usually lower than one per cent of the length ol'the text (Morton 1986:1).

(4) word length [*chrs/wd*]

The average length of the words in a text is another suitable index of vocabulary richness. This feature is calculated by dividing the overall number of (orthographic) characters by the length of the text in words. Word length has been for a long time a lexical characteristic widely used in stylonietric studies (Holmes 1998:113). Zipf (1933) has shown that an inverted relation holds between word length and word Srequency in texts. As Biber (1988:238) has pointed out. "longer words also convey more specific. specialized meanings than shorter ones".

(B) SENTENCE LENGTH

(5) sentence length in characters [*chrs/sent*]
(6) sentence length in words [*wd/sent*]

Despite its limitations (Holmes 1994:89). sentencr length. measured in characters or words. has been used extensively in stylometric works on authorship. As an indicator of style. it is related to the distinction between oral and written registers.

(C) READABILITY GRADES

(7) Automated Readability Index [*arindex*]
(8) Coleman-Liau Index [*clindex*]

Readability evaluation studies. which were born in the 1920s in thr USA. are based on the statistical regularities shown by certain textual linguistic features in relation to their degree of reading comprehension. One of the aims of these studies is the elaboration of empirically tested 'readability formulas' capable of predicting a reading-ease score for a test from a set of selected linguistic features (Gómez Guinovart 1999:83-90). Differences among readability formulas are due to the Iieterogeneity of their derivation from experiments with different texts and subject groups (Klare 1963:33-36). Roth Autoniated Readability Index and Coleman-Liau Index can be easily autoinated (Coleman & Liau 1975:283. Bell Laboratories 1983:156). the formulas applied in our analysis being as follows:

Automated Readability Index = *4.71*letters_per_word + .5*words_per_sentencc – 2*1.43*
Coleman-Liau Index = 5.89*letters_per_word – .3*sentences_per_100_words – 1*5.8*

(D) LEXICAL CLASSES

(9) downtoners [*downts*]: *almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat*

(10) amplifiers [*amplfs*]: *absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very*

(11) discourse particles [*discrs*]: *well, now, anyway, anyhow, anyways*

Our treatment of downtoners, ampliliers and discourse particles draws on Biber's (1988) guidelines. Downtoners and amplifiers constitute two semantic classes of adjuncts. Amplifiers scale the force of the verb upwards from an assumed norm, whereas downtoners have a general lnwering effect on the force of the verb (Quirk et al. 1985:429-430). Roth lists were compiled by avoiding multiwords and items with other major functions (homographs). Discourse particles "are very generalized in their functions and rare outside the conversational genres" (Biber 1988:241). The search for discourse particles in our corpus was limited to sentence-initial occurrences: put differently, only the discourse particles which immediately lollow the BNC new-sentence tag <s> have been considered.

(12) interjections [*interj*]

The grammatical coding of the BNC marks interjections with the attribute ITJ of the part-ol-speech tag <c>. lnterjections indicate emotive aspects of language; they act as discourse markers and thus are characteristic of spoken registers (Schiffrin 1987:73-101). The productivity of interjections in a text is considered a measure ofthe involvement of the speaker with the subject or situation of discourse.

(E) PLACE AND TIME ADVERBIALS

(13) place adverbials [*placeadv*]: *aboard, above, abroad, across, ahead, alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, east, far, hereabouts, indoors, inland, inshore, inside, locally, near, nearby, north, nowhere, outdoors, overland, overseas, south, underfoot, underground, underneath, uphill, upstairs, upstream, west*

(14) time adverbials [*timeadv*]: *afterwards, again, earlier, eventually, formerly, immediately, initially, instantly, late, lately, momentarily, now, nowadays, once, originally,*

presently, previously, recently, shortly, simultaneously, soon, subsequently, today, tomorrow, tonight, yesterday

The massive occurrence of place and time adverbials is taken by Biber (1988:224) as a measure of "situated, as opposed to abstract, textual content". The above lists were compiled from Quirk et al. (1985:516 y 530-531); items with other major functions (homographs) have not been considered.

(F) COORDINATION AND SUBORDINATION

(15) concessive adverbial subordinators [*concessive*]: *although, though*
(16) causative adverbial subordinators [*causative*]: *because*
(17) conditional adverbial subordinators [*conditional*]: *if, unless*

Our analysis of'concessive, causative and conditional adverbial subordinators draws on Biber (1988). According to Biber, these adverbial clauses are found in speech more often than in writing. Riber's list of causative adverbial subordinators excludes lexical forms with other functions (homographs) such as *as, for* and *since*.

(18 j relatives and interrogatives [*wh*]

This group includes *wh*-adverbs (e.g. *when, how, why*), *wh*-determiners (e.g. *which, what, whose, which*) and *wh*-pronouns (e.g. *who, whoever, whom*) either as interrogative or relative introducers (i.e. words marked with the attribute AVQ or DTQ within the BNC) and the word *that* when it introduces a relative clause (attribute CJT). We have not pursued more subtle subclassilication here due to the difficulty in distinguishing automatically between relatives and interrogatives by way of the BNC word-class codes. In further investigation we shall distinguish between *that* and *wh*-words, as in Biber's analysis.

(19) *to*-infinitive clauses [*tocls*]

Infinitive clauses can be identified in the RNC by way of the attribute TO0 in the part-of-speech tag. The output of such automated search includes every *to*-infinitive construction. Unfortunately, it does not allow further discrimination by syntactic function, which Biber's (1988:232) parser does.

(30) coordinating conjunctions [*coord*]

Coordinating conjunctions are coded in the BNC by means of the attribute CJC in the part-of-speech tag. They can be considered indicators of the syntactic complexity of a text. The reliability of coordinating conjunctions as quantitative indicators of style is not absolute since the BNC tagging does not mahe a distinction between phrasal and clausal coordination. each syntactic strategy fulfilling different functions (Biber 1988:245).

(G) VERBAL FORMS

(21) infinitive forms [*inf*]
(22) past participles [*pastpart*]
(33) -*ing* fornis [*ing*]

Infinitive fornis (independent. part of a periphrastic verbal form or preceded by the particle *to* in a *to*-clause). past participles and -*ing* forms are identified in the BNC by way of their word-class codes (VBI. VDI. VHI or VVI for infinitives; VBN. VDN. VHN or VVN for past participles: and VBG. VDG. VHG or VVG for -*ing* fornis). Even though none of them can be unambiguously assigned to a single function. -*ing* forms usually mark progressive aspect. wliereas past participles substantiate perfective aspect or passive voice. All of them have been used as quantitative indicators of style (Moerk 1970:226. McMenamin 1993:195).

(H) MODALS

(24) possibility modal verbs [*posmdl*]
(25) necessity modal verbs [*necmdl*]
(26) predictive modal verbs [*pdtmdl*]
(27) modal auxiliary verbs [*allmdl*]

Modals are marked in the BNC with the attribute VMO in the part-of-speech tag. Following Biber (1988:241). we analyse possibility. necessity and predictive modals separately. In the RNC. possibility niodals include the forms *can. ca. may. might* and *could*; necessity modals include *ought. should* and *must*; and predictive niodals include *will. wo. 'll. would. 'd. shall* and *sha.* The special fornis *ca. wo* and ***shtr*** in the BNC are the segments corresponding to niodals in the negative contractions *can't. won't* and *shan't.* The general category 'modal auxiliary verbs' includes possibility. necessity and predictive modals. along witli other modals not included in Biber's lists such as *used to*. ***dure*** or *need*. which are tagged indistinctly in the BNC.

(I) SYNTACTIC CONSTRUCTIONS

(28) existential constructions [*existential*]

The attribute EX0 in the RNC tags the word *there* appearing in the existential construction *there is*, *there are*, etc. As Riber (1988:228) points out, the stylistic value of this feature is not quite clear since it may be considered an indicator of either "the static, informational style common in writing" or a inarker of "non-complex constructions nith a reduced informational load, (..) more characteristic of spoken registers".

(29) negative constructions [*negative*]

According to Biber (1988:245), "there is twice as much negation in speech as in writing". In this study we have analysed analytic negation (with the negative particle *not* or *n't*), marked in the BNC with the attribute XX0 in the part-of-speech tag.

(30) preposition stranding [*pstranding*]

Stranded prepositions are in some syntactic contexts the unmarked altemative to other more formal pied-piped constructions (Quirk et al. 1985:664). In consequence, the distribution of this feature in the texts is related to the formal versus colloquial dimension of the register under investigation. For the sake of the automation of the analysis, we have only computed stranded prepositions followed by a sentential punctuation mark (exclaination mark, full stop, question mark, colon and seinicolon).

(31) split infinitives [*splitinf*]

The split infinitive is a construction which has been condemned traditionally by prescriptivists of 'good English' (see, for instance, Fowler & Fowler 1922:319) in spite of the fact that in some cases it constitutes the most common alternative in English (Quirk et al. 1985:497). Its absence thus indicates a purist attitude in matters of usage in formal written and oral registers.

(J) PRONOUNS

(32) personal pronouns [*perspron*]
(33) reflexive pronouns [*refxpron*]
(34) indelinite pronouns [*indefpron*]

Personal, reflexive and indefinite pronouns are respectively identified in the BNC by way of the attributes PNP, PNX and PNI in their part-of-speech tag. Personal and reflexive

pronouns have frequently been used as markers of style. Indefinite pronouns. which have been devoted less attention in the literature than personal pronouns. are considered indexes of conceptual abstraction or generalised relerence in the texts (Riber 1988:226).

(K) PREPOSITIONS. ADJECTIVES AND ADVERBS

(35) prepositional phrases [*pp*]
(36) preposition/noun ratio [*p/n*]

Prepositions can be identified in the BNC by means of the attributes PRF (for the preposition *of*) or PRP (for the rest of prepositions) in their part-of-speech tag. They constitute "an important device for packing high amounts of information into academic nominal discourse" and are thus characteristic of informational written discourse (Riber 1988:237). The prepositioninoun ratio. which has been studied with stylometric purposes (McMenamin 1993:198). is regarded as an index of notional complexity.

(37) adjectives [*adj*]
(38) attributive adjectives [*attradj*]
(39) adjectives modified by adverbs [*advadj*]
(40) adjective/noun ratio [*adj/n*]

Adjectives are tagged in the BNC by the attributes AJO (positive adjectives). AJC (comparatives) and AJS (superlatives). Attributive adjectives are identiiied by their occurrence before a noun (Biber 1988:237-38). Both the frequency of adjectives modified by adverbs and the adjectiveinoun ratio have also been used in stylometric studies (McMenamin 1993:196-98). in which they are viewed as indicators of the degree ol'the 'descriptiveness' of a text.

(41) adverbs [*adv*]
(42)adverbipreposition ratio [*adv/p*]

Adverbs –excluding *wh*-adverbs. which belong in feature 18. and including (contrary to Biber 1988:238) downtoners. amplifiers. discourse particles and time and place adverbs. also in features 9 to 1. 13 and 14– are identified in the BNC by the attributes AV0 (general adverbs) and AVP (prepositional adverbs). The adverbipreposition ratio. used in stylometric studies (McMenamin 1993:198). is also related to the degree of 'descriptiveness' of a text.

(L) NOMINAL FEATL **KES**

(43) proper nouns [*propern*]

(44) nouns [n]

Proper nouns are computed in the BNC thanks to the attribute NPO. The role of proper nouns as style indicators is to indicate the degree of concreteness of a text (Julliard 1990). Common nouns are marked in BNC with the attributes NNO. NN1 or NN2. according to their grammatical number. As for their stylistic value. Biber (1988:227) claims that "a high nominal content in a test indicates a high (abstract) informational focus. as opposed to primarily interpersonal or narrative foci".

(45) determiners [*det*]

This group includes words tagged in BNC with the attributes ATO (articles such as *the*. *a*. *an* and no). DPS (possessive determiner forms. e.g. *your*. *their*. *his*) and DTO (general determiners such as *this* or *both*. not tagged as *wh*-determiners. and thus out of the scope of feature 18). The overall stylistic status of this group is associated with textual deixis.

(46) genitive markers [*genitive*]
(47) prepositions plus nouns [*p+n*]
(48) nominal premodiliers [*npremod*]

Nominal categories with accompanying genitive markers are identified in the RNC by the attribute POS. From the perspective of register variation. possessive constructions and nouns preceded by prepositions (both used by Moerk 1970 as marks of authorship). as well as nominal premodifiers (identified in the BNC by the sequence determiner + noun + noun). imply nominal phrase complexity and hence notional depth in the discourse.

III.2. The statistical process

Once the features have been properly explained. in this section we embark on the analysis of each of the statistical steps involved in Biber's multidimensional approach. In what follows. not only shall we describe the techniques but also we will justify the theoretical consequences which each operation has for the methodology as a whole.

The raw countings are normalised to a text length of 1.000 words. which somehow permits the determination of the importance of the features on an intuitive basis. In this connection. an initial reinark seems in order here: if the only purpose of this investigation were the demonstration that there esist enough differences between spoken and written-to-be spoken texts. the comparison of the normalised frequencies of the linguistic features or even their mean values would explain such contrast straightforwardly. As alreadg pointed out. the

main goal of this study is not the corroboration of such difference but the implications which it brings about as far the stylistic characterisation of the text samples is concerned. As Biber (1994:35) points out. "it seems unlikely that the relative distribution of common linguistic features could reliably distinguish aniong registers. In fact. individual linguistic features do not provide the basis for such distinctions".

Table 2 shows the mean values of the features discussed in the previous subsection:

Table 2: Means

	spoken	written
ttr	6.3509522	7.87063
hleg	86.266046	138.8126
hdis	28.483419	37.63046
chrs/wd	4.0268624	4.692527
chrs/sent	116.65821	63.61577
wd/sent	28.839759	13.53329
arindex	11.956404	7.438447
clindex	6.0129306	9.477794
downts	1.630379	1.771187
amplfs	3.5259081	1.481049
discrs	1.8788057	0.529872
interj	13.371244	0.724193
placeadv	2.0459894	3.416971
timeadv	3.8827524	4.578576
concessive	0.3944133	0.422102
causative	2.6544656	0.951116
conditional	4.8488347	1.586188
wh	40.41289	21.59303
tocls	16.971675	16.48863
coord	40.977307	30.00876
inf	41.163607	32.50148
pastpart	18.883174	26.79393
ing	13.730514	15.16303
posmdl	7.2857084	5.648793

	spoken	written
necmdl	1.3813369	1.187515
pdtmdl	8.4167306	6.657576
allmdl	18.626658	13.66377
existential	4.4812804	1.177869
negative	12.237535	5.157774
pstranding	0.5689215	0.190164
splitinf	0.1458038	0.027541
prrspron	99.294638	38.15077
refxpron	1.2340633	0.744659
indefpron	6.0604839	2.378568
pp	84.054743	107.518
p/n	0.5181181	0.383386
adj	42.069145	58.09083
attradj	24.441582	40.83
advadj	5.232514	4.132665
adj/n	11.2614735	0.206409
adv	73.672484	45.33760
adv/p	0.9302901	0.427663
propern	14.136702	63 77199
n	146.81548	218.137
det	115.24982	119.457
genitive	1 936887	5.402313
p+n	18.015697	36.67384
npremod	5.5363913	9969567

The data in this table clearly suggests that sharp differences between spoken and writtení-to-be-spoken) texts can be determined in the light of the mean values of many of the linguistic features investigated: to cite a few. sentence length. frequency of interjections. distribution of attributive adjectives. etc. (The basic descriptive statistics –mean. niinimum. maximum and standard deviation– for each feature are given in Appendix 2.)

The normalised frequencies per 1.000 words are grouped according to their tendency to co-occur. The application of factor analysis to the normalised frequencies leads to the determinatioii of iactors or groups of features which tend with a certain degree of probability either to occur or to be excluded in the texts investigated. In our study case. factor analysis revealed eleven possible factors. as shown in the summary shown in Table 3:

Table 3: Results of factor analysis

Factor	Eigenvalue	% of shared variance	cumulative %
1	13.466	28.1	28.1
2	5.422	11.3	39.4
3	3.828	8.0	47.3
4	3.164	6.6	53.9
5	2.296	4.8	58.7
6	2.116	4.4	63.1
7	1.724	3.6	66.7
8	1.49	3.1	69.8
9	1.29	2.7	72.5
10	1.138	2.4	74.9
11	1.052	2.2	77.1

In the light of the cumulative percentage of shared variance, we shall, in principle, consider four factors (1 to 4), since they constitute more than 50 per cent of the overall shared variance. As Biber (1995b:121) points out, "extracting too many factors is better than too few".[6]

The individual values for the linguistic features under examination are rotated. The statistical technique of rotation allows the discrimination –and subsequent elimination– of those features which are not significant in each factor. The rotation schema which we have used is Varimax rotation, whose main goal is the minimisation of the number of variables on each factor in an attempt to simplify their interpretation. In Table 4 we outline the features which proved relevant for the factors identified by factor analysis, that is, features whose (absolute) factor loadings are bigger than 0.35 (Biber 1988). The order of the features follows, on the one hand, their factor loadings and, on the other, their polarity: whereas positive features are abundant in the factor, the lack of productivity of those features with negative polarity is a defining characteristic of the factor under analysis.

The interest of the information displayed in the previous tables is twofold. First, the ordered list of features in each factor is important with respect to the final interpretation of the factor. In fact, this is the only use of the individual factor loadings. Second, the polarity of each feature within each factor will be decisive as far as the calculation of the factor score of each text.

Once we have got to know which features are positive and which are negative, we associate the normalised frequency of each feature in each text with its corresponding polarity. More specifically, the frequencies for features whose factor loadings are positive will be regarded as positive, whereas the frequencies for features with negative factor loadings will be given negative polarity.

[6] These four factors contain a considerable number of factor loadings bigger than 0.35, which indicates that their selection out of the eleven factors recognised by factor analysis is in the driving seat.

Table 4: Factor loadings of the significant features

features>.35	Factor 1	
clindex	0.95006	+
chrs/wd	0.94236	+
attradj	0.90835	+
n	0.90595	+
p+n	0.86004	+
pp	0.84553	+
adj	0.81473	+
propern	0.68458	+
npremod	0.55403	+
det	0.50225	+
pastpart	0.46815	+
genitive	0.39371	+
hleg	0.37412	+
wh	-0.37011	-
inf	-0.38944	-
coord	-0.50238	-
discrs	-0.50537	-
p/n	-0.53134	-
negative	-0.59465	-
adv	-0.60609	-
interj	-0.66444	-
indefpron	-0.7199	-
adv/p	-0.77787	-
perspron	-0.91324	-

features>.35	Factor 2	
posmdl	0.73169	+
wh	0.71153	+
advadj	0.69877	+
adj/n	0.68769	+
amplfs	0.66988	+
allmdls	0.65887	+
conditional	0.59404	+
causative	0.56265	+
inf	0.55157	+
pdtmdl	0.48861	+
adv	0.46532	+
necmdl	0.45066	+
ttr	-0.36403	-
placeadv	-0.36894	-
propern	-0.51689	-

features>.35	Factor 3	
wd/sent	0.89642	+
chrs/sent	0.894	+
arindex	0.89154	+
refxpron	0.43924	+
ttr	-0.56899	-

features>.35	Factor 4	
pstranding	0.71538	+
bdis	0.65769	+
genitive	0.60933	+
tocls	0.59019	+
npremod	0.44046	+
conditional	0.4243	+
discrs	0.40906	+
timeadv	0.39999	+
inf	0.39816	+
hleg	0.36885	+

The following step is the standardisation of all the normalised frequencies according to the same criterion. In this case. following Biber. we have used the comnion standardisation criterion to a mean of 0.0 and a standard deviation of 1.0. In more detail. the normalised frequency of the feature in a given text niinus the mean frequency of the feature in all the samples investigated is divided by the standard deviation of the feature. The standardised result per feature per text is called the 'standardised score' of the feature. The score for the whole factor is coniputed by adding the standardised scores of the features with positive factor loadings and by subtracting the standardised scores of those with negative loadings. The score for a group of tests (text type or genre) is done by calculating the mean of the factor scores for all the texts which belong to the same sub-/genre. To illustrate the whole process. in Appendis 3. we show the standardised scores of the features corresponding to factor 3 as well as the factor scores of the genres investigated. Table 5 gives the results for the whole corpus according to the four factors identified by factor analysis:

Table 5: Summary of the scores per text type

text type	text-type divisions	Factor 1	Factor 2	Factor 3	Factor 4
written	prayers	12.2098373	-10.5612284	2.20912148	-4.35518811
	sermons	10.5677101	-4.36446306	-0.00303066	-2.14445519
	speeches	15.0714213	-0.12745265	-0.63279867	1.39969937
	TV news	23.3430073	-13.6323543	-1.83665774	-0.53257161
	applied science	29.0171926	-5.33478016	-0.35330197	4.47519478
Total		24.3074677	-10.7341919	-1.26828537	0.82674472
spoken	responses	-10.3825078	6.35289683	2.88881601	-2.46083851
	lectures	-2.91097664	7.67925377	0.2261068	-0.65477574
	courses, presentations	-9.95310282	6.32303802	-0.62650913	4.74424374
	talks	-10.5615823	5.0251022	-0.84427013	-0.67394113
	news	10.9582372	-3.34438549	0.37030004	1.7984393
	sermons	-6.1757523	0.68559179	7.55585622	-1.67005743
	speeches	3.17090854	1.93386535	0.85328597	0.49105851
	interviews	-15.0671657	-1.05027663	-1.11885211	-1.52516046
	meetings, debates	-0.31334219	-2.84701657	-0.46482087	0.43387244
	court/parliament cases	0.66585214	6.04991021	0.98774435	-0.92197894
Total		-6.92480185	3.05799654	0.36131385	-0.23552611

The results in this table will be used in order to place even text type and text-type division on a factor scale. which will give an idea of the characterisation of each genre along the interpretation of the factor. which constitutes the main topic of the following section.

IV. INTERPRETATION OF THE RESULTS

Biber's multifeature multidimensional methodology has been shown to be a precise way of quantifying the stylistic characterisation of a genre. By means of factor analysis. further rotation techniques and the standardisation of the results. we have achieved unique values for the text types and text-type categories recognised in our corpus. Such quantitative data by itself demonstrates that the linguistic nature of the texts under investigation is clearly different and. on many occasions. divergent. Nonetheless. the backbone of this methodology is the functional or. as Biber puts it. dimensional characterisation of the textual material. To that end. each factor or group of features is assumed to represent a dimension. either bipolar (factors or Dirnensions 1. 2 and 3 in our corpus study) or monopolar (factor o dimension 4). The dimensions to which we shall pay attention in what follows are as follows: (The features with double bracketing have factor loadings from 0.35 to 0.4. which were not considered in. for example. Biber & Finegan 1997: the factor loadings of those with single bracketing go from 0.4 to 0.45. which are disregarded in Kim & Riber 1994.)

DIMENSION 1

Positive features

Coleman-Liau Index [*clindex*]
word length [*chrs/wd*]
attributive adjectives [*attradj*]

nouns [*n*]
preposition plus noun [*p+n*]
prepositional phrases [*pp*]
adjectives [*adj*]
proper iiouiis [*propern*]
nominal premodifiers [*npremod*]
determiners [*det*]
past participles [*pastpart*]
((genitive markers)) [*genitive*]
((*hapax legomena*)) [*hleg*]

Negative features

((relatives and interrogatives)) [*wh*]
((infinitive forms)) [*inf*]
coordinating conjunctions [*coord*]
discourse particles [*discrs*]
prepositioii'noun ratio [*p/n*]
negative constructions [*negative*]
adverbs [*adv*]
interjections [*interj*]
indefinite pronouns [*indefpron*]
adverb/preposition raiio [*adv/p*]
personal pronouns [*perspron*]

DIMENSION 7

Positive features

possibility modal verbs [*posmdl*]
relatives and interrogatives [*wh*]
adjectives modified by adverbs [*adv/adj*]
adjective/noun ratio [*adj/n*]
amplifiers [*amplfs*]
modal auxiliar) verbs [*allmdls*]
conditional adverbial subordiiiators [*conditional*]
causative adverbial subordinators [*causative*]
infinitive foinis [*inf*]
predictive modal verbs [*pdmdl*]
adverbs [*adv*]
(necessity inodal verbs) [*necmdl*]

Negative features

((type-token ratio)) [*ttr*]
((place adverbials)) [*placeadv*]
proper nouns [*propern*]

DIMENSION 3

Positive features

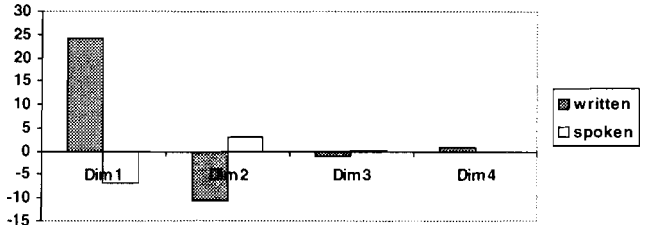
senience leiigtli iii words [*wd/sent*]
sentence length in characters [*chrs sent*]
Autoniated Readability Index [*arindex*]

(reflexive *pronouns)* [*refxpron*]
Negative features
type-token ratio [*ttr*]

DIMENSION 4
Positive features
preposition stranding [*pstranding*]
hapax legomena [*hdis*]
genitive markers [*genitive*]
to-infinitive *clauses* [*tocls*]
(nominal premodifiers) [*npremod*]
(conditional *adverbial* subordinators) [*conditional*]
(discourse particles) [*discrs*]
((time adverbials)) [*timeadv*]
((infinitive forms)) [*inf*]
((*hapax legomena*)) [*hleg*]
Negative features
none

The differences between the spoken and written-to-be-spoken samples of the corpus plotted in Graphic 1 shows that. whereas the first two dimensions are of great importance as far as the overall stylistic explanation of the text type investigated. the contrast evinced by the textual material along Dimensions 3 and 4 is not explanatory enough *per se*. Such a fact will account for the special attention which shall be paid to the interpretation of Dimensions 1 and

Graphic 1: Dimensions per major text categories



The interpretation of the dimensions is the hardest stage in the multidimensional model. Such interpretation acts as a kind of heading which has to account for both the successful features and their polarity within each factor. Examples of dimensions are as follows: 'informational versus involved production'. 'elaborated versus situation-dependent reference'. 'abstract style'. 'narrative versus non-narrative concems' and 'overt expression of persuasion'

(Biber 1988. Biber & Finegan 1989. 1992). 'on-line interaction versus planned exposition'. 'overt versus implicit logical cohesion' and 'overt personal stance' (Kim & Biber 1994). or 'argumentative versus reported representalion of information' (Biber & Hared 1994). In what follows. we shall concentrate on the interpretation of Dimensions 1 and 2.

IV.1. Dimension 1: 'notional richness versus dynamic deictic reference'

The features in this dimension can be grouped into the following categories. according to their polarity: (i) features associated with notional complexity. (ii) lack of dynamic grammatical categories and devices. (iii) lack of personal reference and (iv) lack of vague reference.

The features associated with notional complexity involve. on the one hand. structural complexity. that is. word length (Coleman-Liau index and word length in number of characters) and. on the other. percentage of nominal categories. The amount of nominal categories is shown not only by the productivity of nouns but also by the percentage of attributive adjectives and -ed forms. prepositions followed by nouns. determiners and genitive markers (*'s*). all they implying the existence of nouns heading nominal projections. All these features. together with the importance of proper nouns. share the same objective. namely. an increase in the notional depth of the discourse. That is the reason why we have chosen the heading 'notional richness' for the positive dimension of this factor.

By contrast. the two remaining issues are clearly associated with the negative features grouped under this dimension. More specifically. the low percentages of discourse particles (and also subordinating *wh*-introducers). of independent adverbs in general and of adverbs with respect to prepositions[7] signify that the style is not dynamic but noun-centred and carefully planned. The preference for self-defining nouns and the avoidance of pronouns with either situational (personal pronouns) or indefinite (indelinite pronouns) referents indicates that the discourse is impersonal and precise.

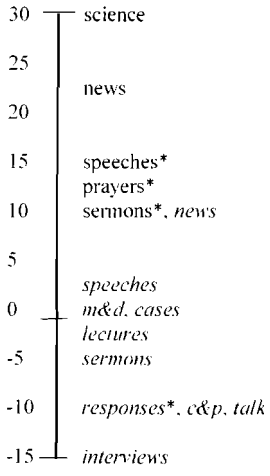
The scale oftext-type categories along Dimension 1. drawing on the statistical data in Appendix 4. is given in Graphic 2 (genres in italics correspond to the spoken texts. whereas those in Roinan case belong to the written-to-be-spoken material: asterisked genres are those whose statistical data is based on only one sample[8]):

---

[7] The negative influence of tlie proportion of adverbs over the grammatical class of prepositions seems to indicate that the frequency ofprepositions. most of which govern nouns within prepositional phrases (the number of stranded prepositions does not even constitute a significant feature in this factor). surpasses that of adverbs. Such a fact ultimately gives support to the importance of nominal categories in the dimension under analysis.

[8] Even though we are aware that mono-sample genre investigation is not allowed within the multidimensional approach since it wreaks havoc with the prerequisite ofcorpus representativeness already discussed. we have used mono-sample text typologies in an attempt to wide the spectrum of the genres under discussion. Needless

Graphic 2: Dimension 1



Several remarks seem in order here in the light of Graphic 2. First. the textual categories investigated are splendidly characterised along the dimension suggested for the first factor: whereas every written text type is associated with a positive factor loading. only two spoken genres are positive. namely. news and speeches. which are indisputably less interactive than the rest of the spoken samples. More specifically. 'orthodox' news and speech productions do not allow feedback from the listener(s). which is possible (and on many occasions implied) in the case of meetingldebates. cases. lectures. sermons. responses. courses/presentations. talks and interviews.

In the previous paragraph we have justified the dichotomy written *vs* spoken which is shown by the results on Dimension 1 in terms of interaction. The degree of interaction associated with a linguistic production is in keeping with other features such as spontaneity and. by extension. dynamic deictic reference. which is part of the label choseii for Dimension 1. On ihe one hand. the importance of the listener in the discourse brings about potential changes on the speaker's part. who cannot stick to a prefixed script and is obliged to admit new elements without a corresponding agenda. Thus. discourse dynamics play a fundamental role in the diniensional characterisation of the text types under discussion. On the other hand. many of the elements iii an interactive production are taken from or simply grounded on the environmental circumstances in which the discourse takes place. This consequence leads us to the second part of the defining headline. namely. the deictic dimension.

The opposite situation is pictured by the wriiien texts. In this respect. the scale sketched

---

10 su). the results obtained by text categories based on only one text will be treated with extreme caution.

in Graphic 2 is revealing. Sermons and prayers (and. possibly. speeches as well) are the text categories which are the most interactive among the group of non-interactive text types. In these genres. even though interaction is not real. it is regarded as a compulsory part of the fictional communicative process. The elaboration of a speech. prayer or sermon involves not only the actual physical presence oí' the listener but also the effect which the linguistic production is going to have on him or her. In an attempt to get the attention of the potential listener(s). the writer ot' sermons. prayers. etc. is likely to introduce some deictic elements in the composition and to make use of thematic variation. changes in the discourse topic being. in consequence. of considerable significance.

IV.2. Dimension *2:* 'explicitness versus concision'

The linguistic features which proved to be significant in Dimension 2 are indicators of either syntactic expansion or lexical richness. In what follows we justify the ascription of the features to the previous categories.

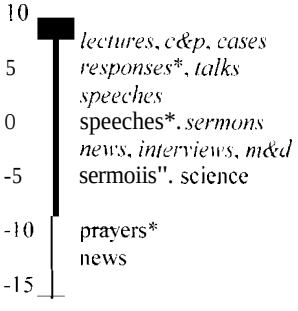
As far as syntactic expansion is concerned. the features which materialise the speaker's attempt to increase the actual length of his or her linguistic productions. either at phrasal or clausal level. by adding elements which do not contribute to the progression of the discourse but simply enlarge the descriptive burden of the active referents. can be grouped into two categories: features associated with syntactic clausal modification or subordination. and features related to phrasal modification. The former are substantiated by subordinating sentence-introducers and conditionalicausative adverbial subordinators. The features implying syntactic modification at clause level are: modals (including the generic label modal plus possibility. prediction and necessity modal auxiliaries). adverbs premodifying adjectives. adjectiveinoun ratio. amplifiers and general adverbs. In more detail. modals modify verbal groups; adverbs modify adjectives: most adjectives are attributive and thus act as nominal premodifiers: amplifiers are modifying categories by definition and. finally. general (non-*wh*) adverbs are used to modify either adjectives/adverbs (degree adverbs) or the whole predication/proposition (a limited number of adverbs are complements in the sentences in which their occur). We have excluded the feature 'infinitive forms' from the previous description because such nonfinite forms are twofold in that they can serve as dependents of either nominal constructions (*an attempt* [*to* he *more* precise]) or of predicators (*I want* [*to be more precise*]). In both cases the· increase the syntactic complexity of the utterance and do not necessarily enlarge the set of the active referents relevant for the progression ofthe discourse.

With respect to the second pole of the interpretation. that is. lexical richness. the only feature which is statistically significant and. thus. has consequences for the interpretation of Dimension 2 is 'proper nouns'. Proper nouns can be understood as the category that introduces referents in the discourse by using the least amount of linguistic material. In this connection.

proper nouns do not require (in standard English. they do not tolerate) further syntactic elaboration by means of determiners. modifying adjectives. etc. Another related feature of less statistical importance is the so-called type/token ratio. The inclusion of type/token ratio. understood as an indicator of lexical diffusion. among the negative group of features corroborates the label given in the title of this section. namely. concision. She less the relevance of the index of lexical variability in a given text or group of texts. the more concise the style of tlie texts." If the number of significant factor is restricted to 3. a new variable joins the group of the negative features. namely. possessive-*'s* constructions."' The inclusion of genitive constructions complies with the functional interpretation of the negative pole. that is. concision. since this type of syntactic strateg): implies both the existence of premodifying nominal categories (*'s* is attached to nominal projections) and the avoidance of other periphrastic ways of materialising possession in English. such as the *of*-construction.

Graphic 3 rellects the relative position of the text categories investigated along Dimension 2:

Graphic 3: Dimension 2



The data in Graphic 3 are revealing as far as the characterisation of texts according to Dimension 3 is concerned. Every written text investigated is located below 0 in the scale. As shown in the graphic. the news material is quite detached from the othei spoken text types. Since most of the journalistic texts included in the corpus are TV news. the style of this genre is almost telegraphic due to timing reasons and thus considerably concise. which keeps track oi'the position the text type occupies in the previous scale. By contrast. most of the spoken material is associated with positive values (lectures. courses/presentations. cases. responses.

[9] The factor loading for tlie type/token ratio becoiiies -.34 if Varimax rotation is applied to only ihree factors. Such a loading would imply tlie disappearance of the 'ttr' feature from the list of significant variables in Dimension 2 (see section 3.4 in this connection).

[10] The loading for tlie feature 'genitive' after the application of Varimax rotation to ihe three-factor daia is -.5. which denioiistrates ttie significance of the feature.

talks and speeches).

Even though explicitness versus concision seems to be relevant for the distinction between spoken and written-to-be-spoken linguistic productions. the differences among the samples do not allow statistical validation. In this connection. the contrast holding between the spoken text type with the highest (positive) value in Graphic 3. that is. lectures (7.67) and the written genre with tlie lowest (negative) loading. namely. TV news (-33.63) is less prominent than that between the poles within the group of spoken texts in Graphic 2 (Diniension 3).

IV.3. Dimensions 3 and 4

The data corresponding to Dimensions 3 and 4 must be treated with caution since the differences among the texts analysed make it hard to assess the suitability of the application of these diniensions to tlie stylistic characterisation of spoken/written-to-be-spoken texts. Graphic 1 above showed that the mean values for the two major classes are actually similar along the dimensions under discussion.

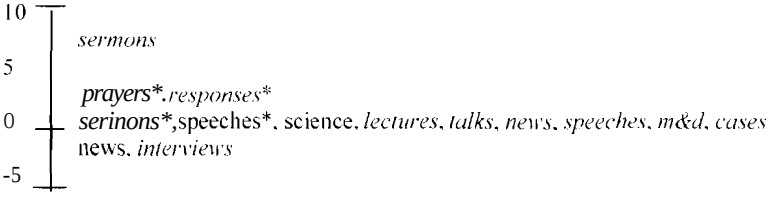
The significant linguistic features relevant to Dimension 3. which could be interpreted in terms of 'favoured versus disfavoured (sentential) length'. are scarcely two. namely. (positive) sentence length (word length. character length and ARI) and (negative) type-token ratio (lexical variability). The inclusion of reflexive pronouns in the group of the positive features. which is not totally significant from a statistical point of view. can be explained by way of the inclusion of every *self/selves*-form – eitlier reflexive. that is. argumenta1 or eniphatic. that is. syntactically unnecessary. Whereas reflexive pronouns cast doubt on tlie final interpretation of Dimension 3. the coiisideration of emphatic *self/selves*-forms. which are syntactically and semantically optional. countenance the functional heading of '(sentential) length'.

A new situation will emerge as far as Dimension 3 is concerned if tlie fourth factor is discarded since new linguistic features will enter into the picture. On the one hand. the preposition/noun ratio. coordination and the existential construction will get significant values after rotation (+.5. +.44 and +.4. respectively). On the other. the rotation of only three factors will lead to the inclusion of *hapax dislegomena*. nominal premodification and preposition stranding in tlie set of negative features. Such new features are in keeping with the interpretation already discussed. First. the increase of the proportion of prepositions against nouns illustrates the productivity of prepositional modifiers andior complements and. thus. the tendency to long utterances. By contrast. the significance of nominal premodification. that is. nouns premodifying other nouns. which. according to the three-factor analysis have to be placed in the negative group of features. implies the preference for short syntactic constructions. Second. tlie relevance of sentence length to the functional characterisation of the texts along Diniension 3 is emphasised by the high frequencies for coordination – either

clausal or phrasal. Third. the /here-existential construction. understood as a syntactic device for estraposing the subject of tlie sentence to postverbal position. requires tlie insertion of dummy there in the canonical preverbal subject position. Needless to say. the duplication of syntactic constituents increases the length of the sentence (There is a unicorn in the garden versus A unicorn is in the garden). Fourth. preposition stranding can be seen as a way of favouring covert relative pronouns in sentences with embedded relative clauses. Put differently. relative pronouns are compulsory in pied-piped constructions. that is. in relative clauses in which the prepositions govern the relative pronouns (This is the paper to {which / *Ø} I devoted my life). whereas they can be omitted if the preposition is left stranded in postverbal position (This is the paper {which / Ø} I devoted my life to). From this perspective. tlie appearance of the feature preposition stranding on the negative side of Dimension 3 highlights the preference Sor short utterances. Finally. the type/token ratio and hapax dislegomena deserve our attention in the discussion of the features which are significant in Dimension 3 – the statistical importance of the former is granted by both three- and four-factor analysis. whereas the latter proved to be significant only on a three-factor basis. Both features are indicators of lexical variability and meet proper explanation here since. as already pointed out. lexical richness correlates with syntactic niinimisation.

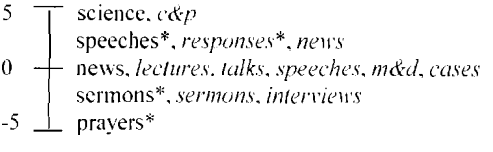
Graphic 4 shows the scale portraying the location of the text types along Dimension 3:

Graphic 4: Dimension 3



The relevance of Dimension 4 to the functional characterisation of spoken texts collapses upon tlie examination of the results reflected in Graphics 1 above and 5 below.

Graphic 5: Dimension 4



The differences among the samples are even more subtle than they were in the case of Dimension 3. Spoken and written-to-be-spoken texts occupy the central positions (from 4 to -

4). To our knowledge. since no plausible hunch seenis to rule such a situation. 'syntactic/lexical markedness' cannot be clainied to exert any influence on the characterisation of the corpus material. The heading suggested here attempts the rather pointless business of labelling the functional consequences which the linguistic features in Dimension 4 have. Whereas syntactic markedness is represented by preposition stranding. genitive markers and noun phrases consisting of deterniiners followed by two nouns. lexical markedness would be implied by *hapax legomena* and *dislegomena*.

The typologies already described seeni to point towards the direction that a functional characterisation of spoken and written-to-be-spoken textual material by means of the four dimensions discussed above does not stand a chance of survival. A further avenue is thus open to us. namely. the reduction of the nuniber of operative factors to three. whose consequences have already been outlined in the description of Dimensions 1 to 3.

V. CONCLUSIONS AND FINAL REMARKS

In this paper we have applied Biber's multifeature niultidimensional analysis to a corpus of 3.000.000 words taken froni the British National Corpus. comprising spoken and written-to-be-spoken texts in an attenipt to demonstrate. first. that these two major textual categories differ considerably in style and. second. that their functional characterisation can be achieved by the observation and the subsequent statistical treatnient of the results on the occurrence of carefully-designed sets of significant linguistic features.

The application of factor analysis led to the distinctioii of four factors. two of which kept indisputable track of the contrast which the text types investigated (written: prayers. sernions. speeches. TV news. applied science: spoken: responses. lectures. courses/presentations. talks. news. sermons. speeches. interviews. nieetingsidebates. courtiparlianient cases) showed when they were located on the diniensional scales resulting froni factor analysis. It was precisely such relative scaling that cast doubt on the significance of the two reniaining factors.The bulk of the second half of the paper was taken up by the analysis of the adequacy of the linguistic features and the textual samples themselves with respect to the two factors which the multidimensional model revealed were relevant to the characterisation of spoken and written-to-be-spoken texts. In a nutshell. tlie data confirmed that such two major textual categories can be distinguished as follows: on the one Iiand. whereas spoken texts tend to incorporate dynamic deictic relerents. written texts favour tlie occurrence of elaborate concepts: on tlie other hand. the spoken texts were considerably more explicit than the written-to-be-spoken saniples investigated. The interpretations corresponding to the two other factors which did not shown to be so significant since they substantiated niinor differences aniong the texts were. respectively. 'favoured versus dislavoured (sentential) length' and 'syntactic/lexical markedness'.

In further investigation, we shall try, first, following Biber (1994), to make use of situational parametres such as mode, interaction, careful production, informative purpose, etc. and correlate them with the linguistic features. Second, both the selection and the definition of the linguistic features will undergo some changes. On the one hand, the number of significant linguistic variables can be enlarged by the introduction of additional features (e.g. nominalisations, as in Biber 1988). On the other, several of the features computed in this pilot study may require further subclassification (e.g. phrasal/clausal coordination, verbal tenses, 1st/2nd/3rd person personal pronouns, etc.) and refinement.

REFERENCES

Aston. G. & L. Rurnard (1997) *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh U.P.

Atkinson. D. (1992) "The Evolution of Medical Research Writing from 1735 to 1985: The Case of the *Edinburgh Medical Journal*". *Applied Linguistics*. 13(4): 337-74.

Rarnbrook. G. (1996) *Language and Computers A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh U.P.

Bell Laboratories (1983) *Unix Writers Workbench Software User's Manual*. Piscataway: Bell Laboratories.

Besnier. Niko (1988) "The Linguistic Relationship of Spoken and Written Nukulaelae Registers". *Language*. 64: 707-36.

Riber. D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge U.P.

– (1994) "An Analytical Framework for Register Studies" in Biber & Finegan (eds.). 31-56.

– (1995a) "On the Role of Computational. Statistical. and Interpretive Techniques in a Multi-dimensional Analysis of Register Variation. A Reply to Watson". *Text*. 15(3): 341-70.

– (1995b). *Dimensions of Register Variation. A Cross-linguistic Comparison*. Cambridge: Cambridge U.P.

Riber. D. & E. Finegan (1989) "Drift and Evolution of English Style: A History of Three Genres". *Language* 65(3): 487-517.

– (1992) "The Linguistic Evolution of Five Written and Speech-Based English Genres from the 17th to the 20th Centuries" in M. Rissanen. O. Ihalainen. T. Nevalainen & I. Taavitsainen (eds.) *History of Englishes: New Methods and Interpretations in Historical Linguistics*. Berlin: Mouton de Gruyter. 688-704.

– (1997) "Diachronic Relations among Speech-Based and Written Registers in English" in T. Nevalainen & I. Taavitsainen (eds.) *To Explain the Present. Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique. 253-75.

Biber. D. & E. Finegan (eds.) (1994) *Sociolinguistic Perspectives on Register*. New York: Oxford U.P.

Ribei. D. & M. Hared (1994) "Linguistic Correlates of the Transition to Literacy in Somali: Language Adaptation in Six Press Registers" in Biber & Finegan (eds.). 182-216.

Burnard. L. (199.5) *User Reference Guide for the British National Corpus*. Oxford: Oxford University Computing Services.

Coleman. M. & T.L. Liau (197.5) "A Computer Readability Formula Designed for Machine Scoring". *Journal of Applied Psychology*. 60(2): 283-284.

Fowler. H.W. & F.G. Fowler (1922) *The King's English*. Oxford: Clarendon.

Gómez Guinovart. J. (1999) *La escritura asistida por ordenador: problemas de sintaxis y de estilo*. Vigo: Universidade de Vigo (Servicio de Publicacións).

González-Álvarez. D. & J. Pérez-Guerra (1997) "A Corpus-Based Approach to the Multidimensional Variation of Four Late Middle English Genres". paper read at the *10th International SELIM Conference*. Universidad de Zaragoza.

– (1998) "Texting the Written Evidence. On Register Analysis in Late Middle English. *Text*. 18(3): 321-48.

Holmes. D.I. (1994) "Authorship Attribution". *Computers and the Humanities*. 28: 87-106.

– (1998) "The Evolution of Stylometry in Humanities Scholarship". *Literary and Linguistic Computing*. 13(3): 111-117.

Julliard. M. (1990) "Proper Nouns as Proper Style-Markers of Poetry and Prose". *Literary and Linguistic Computing*. 5(1): 1-8.

Kim. Y-J. & D. Biber (1994) "A Corpus-Rased Analysis of Register Variation in Korean" in Biber & Finegan (eds.). 157-81.

Klare. G.R. (1963) *The Measurement of Readability*. Ames: Iowa State U.P.

McMenamin. G.R. (1993) *Forensic Stylistics*. Amsterdam: Elsevier.

Moerk. E.L. (1970) "Quantitative Analysis of Writing Styles". *Journal of Linguistics*. 6: 223

Morton. A.Q. (1986) "Once. A Test of Authorship Rased on Words Which Are Not Repeated in the Sample". *Literary and Linguistic Computing*. 1(1): 1-8.

Quirk. R.. S. Greenbaum. G. Leech & J. Svartvik (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Schiffrin. D. (1987) *Discourse Markers*. Cambridge: Cambridge U.P.

Zipf. G.K. (1932) *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge: Harvard U.P.

APPENDICES

Appendix 1: AWK program for lexical types in BNC

```
{
for (i=1; i<=NF; i++)
      if ($i ~ /<w/)
      {
      start = index($(i+1),">")+1
      if (index(substr($(i+1),start),"<"))

            Words[tolower(substr($(i+1),start,
            length(substr($(i+1),start))-2))]++
            }
      else
            {
            Words[tolower(substr($(i+1),start))]++
            }


END {
      for (w in Words)

                  diff++
                  }
      printf(" i", diff)
```

Appendix 2: Descriptive statistics

	Mean	StdDev	Minimum	Maximum
ttr	6.60	3.13	0.14	17.27
hleg	97.92	62.32	9.76	379.19
hdis	30.51	16.34	2.92	168.67
chrs'wd	4.17	0.39	3.51	4.99
chrs'sent	104.9	235.74	20.76	2866.67
wd/sent	25.45	59.51	5.79	721.33
arindex	10.95	29.71	-1.65	357.95
clindex	6.78	2.62	0.14	11.8
downts	1.66	0.75	0	4.11
amplfs	3.07	2.1	0	12.23
discrs	1.58	1.48	0	12.05
interj	10.57	16.15	0	107.21
placeadv	2.35	1.47	0	8.01
timeadv	4.04	1.95	0	14.42
concessive	0.4	0.42	0	2.97
causative	7.28	1.61	0	7.3
conditional	1.13	3.49	0	24.1
wh	36.25	12.02	1.32	72.29
tocls	16.86	5.44	0.49	53.13
coord	38.55	10.23	12.05	74.31
inf	39.24	9.22	3.17	71.82
pastpart	20.64	5.91	6.67	42.68
ing	14.05	3.66	4.73	25.34
posmdl	6.92	2.97	0	15.38
necmdl	1.34	1.03	0	7.54

pdtmdl	8.03	3.18	0	18.52
allmdl	17.53	6.09	0	50.86
existential	3.97	2.12	0	13.27
negative	10.67	5.89	0	38.58
pstranding	0.48	0.96	0	12.05
splitinf	0.12	0.2	0	1.2
perspron	85.74	38.06	16.11	161.12
refxpron	1.13	0.96	0	7.04
indefpron	5.74	3.39	0	24.1
PP	89.76	18.02	4	151.34
p/n	0.5	0.09	0.26	0.82
adj	45.62	15.17	0	97.48
attradj	4.99	2.63	0	14.6
advadj	28.08	12.1	0	60.71
adj/n	0.25	0.07	0	0.71
adv	67.39	18.39	19.34	137.63
adv/p	0.82	0.36	0.13	2.69
propern	25.14	24.13	0	179.53
n	162.63	39.96	62.06	250.96
det	116.18	13.89	56.38	156.45
genitive	7.71	2.56	0	24.1
p-n	72.15	9.88	0	50
npremod	6.57	3.64	0	14.1

Appendix 3: Scores for written texts in factor 3

text (genre)	wd/sent	chrs/sent	arindex	refxpron	ttr	text factorial	text-type factorial	genre score
GX0 (prayers)	-0.11942079	-0.11575416	-0.11049739	2.16018705	-0.39460677	2.20912148	2.20912148	
GX1 (sermons)	-0.1512594	-0.15409962	-0.15223041	0.2606739	-0.19388488	-0.00303066	-0.00303066	
JH (speech)	-0.130489	-0.1044325	-0.07371485	-0.65336934	-0.32970702	-0.63279867	-0.63279867	
JHM (TV news)	-0.23718871	-0.2092292	-0.12179213	-0.78349965	0.68291304	-2.03462275		
JHN	-0.19777897	-0.16461824	-0.0037148	2.09182499	0.19938768	1.43628662		
K20	-0.23916555	-0.22027678	-0.15243483	-0.56594907	0.71249416	-1.89032039		
h21	-0.24994868	-0.23637889	-0.17470056	-0.28630303	0.88544328	-1.83277444		
K22	-0.24770522	-0.23361311	-0.17100471	-0.45411711	0.84775256	-1.96038293		
h 2	-0.24522068	-0.23005911	-0.16732049	-0.4033425	0.80709385	-1.85303664		
h24	-0.24172506	-0.22877810	-0.16746809	-0.57938161	0.78114824	-2.0025014		
K25	-0.24896703	-0.23581988	-0.17580455	-0.43777185	0.86883489	-1.96719821		
K26	-0.24109914	-0.22651748	-0.16768932	-0.70011016	0.74203513	-2.08649123		
K27	-0.21418003	-0.21939568	-0.16388446	-0.5555043	0.63905136	-1.81201583		
K28	-0.25856348	-024501614	-0.18125438	-0.74687539	1.03947175	-2.47210135		
K1B	-0.23697389	-0.21987585	-0.15752927	-0.54316193	0.67973519	-1.81727612		
K1C	-0.24988062	-0.23549992	-0.17180866	-0.05343193	0.88428593	-1.59490707		
K1D	-0.24160775	-0.23349348	-0.16559072	-0.69326437	0.87965432	-2.22161064		
K1L	-0.2419015	-0.22581714	-0.16287819	-0.37591293	0.75447407	-1.76098404		
K1F	-0.24323641	-0.226'10798	-0.16256444	-0.54348831	0.77540887	-1.951606		
K1G	-0.24438021	-0.22980754	-0.16901138	-0.5667358	0.79358944	-2.00352439		
K1H	-0.24532542	-0.22892409	-0.16308939	-0.59408299	0.80878547	-2.04020736		
h1J	-0.24021214	-0.22240216	-0.15672572	-0.70879368	0.72840805	-2.05661195		
h1h	-0.24951548	-0.23530134	-0.17226046	-0.53369291	0.87809153	-2.06886172		
K1L	-0.24778717	-0.21372805	-0.17214194	-0.58035642	0.8491128	-2.08312637		
K1M	-0.24970623	-0.23436938	-0.16836849	-0.20193917	0.88132424	-1.73570751		
K1N	-0.24709724	-0.2314448 1	-0.16632705	-0.43234244	0.8376997	-1.91491124		
K1P	-0.2502178	-0.23645379	-0.17411641	-0.66947314	0.89002874	-2.22028989		
K1R	-0.24272032	-0.22691538	-0.16414636	-0.51688747	0.76727939	-1.41794841		
K1S	-0.2377901	-0.2216877	-0.16115256	-0.43939893	0.69184672	-1.75187601		
K1T	-0.24172023	-0.22544412	-0.16214949	-0.41685331	0.75165441	-1.79781862		
K1U	-0.24595487	-0.23014746	-0.16536989	-0.40546425	0.81899288	-1.86592935		
K1V	-0.24583292	-0.2307535	-0.16784006	-0.57055398	0.81700974	-2.0319902		
h1W	-0.23368636	-0.21617259	-0.15481617	-0.78502206	0.63198441	-2.02168159		
K1X	-0.24481304	-0.22963188	-0.16709485	-0.50040776	0.80052847	-1.942476		
K1Y	-0.23428832	-0.22039706	-0.16683588	-0.21639649	0.64060626	-1.47852401	-1.83665774	
K2G (applied science)	-0.1646607	-0.11359945	-0.03522614	-1.172451177	-0.094861h	-1.39107546		
K2P	-0.17676106	-0.15287193	-0.10748948	-0.7804046	0.00363918	-1.22116626		
K2X	-0.12437528	-0.07281008	-0.01555428	-0.31722362	-0.36614132	-0.16182191		
K3G	-0.17123514	-0.14556794	-0.09982658	0.41545573	-004249719	0.04132327		
K3D	-0.68711334	-0.04319714	-0.00773623	-1.17245077	-0.55808488	-0.751 71261		
K3L	-0.09607538	-0.05137807	-11.01241109	0.3361170	-0.51728547	0.69353884		
K3U	-0.11656164	-0.07079848	-0.02303093	-0.11060911	-0.41062106	0.0896209		
K42	-0.10396165	-0.04607001	0.008 15974	-1.17245077	-0.47782319	-0.83649951		
K48	-0.13210555	-0.08944554	-0.03848697	-0.78778804	-0.31982125	-0.72800487		
K4L	-015387254	-0.11167926	-0.05121046	-0.67285045	-0.17533723	-0.81427547		
K4N	-0.08076617	0.02088553	0.02522354	-0.84472805	-0.58876727	-0.33238411		
K4X	-0.10728413	-0.0444977	0.01594411	-0.33 1714	-0.46061598	-0.00694074		
K56	-0.05579998	0.00559327	0.04338818	-0.087028	-0.69271369	0.59886716		
K5L	-0.07069982	-0.00285915	0.04536587	-0.7279273	-0.63242865	-0.12369175	-0.35330197	-1.26828537

Appendix 4: Scores for the dimensions

		Diineiisioii 1	Dimension 2	Diineiisioii 3	Diineiisioii 4
	prayers	12.2098573	- 10.5612284	2.20912148	-4.3551881 1
	seriiions	10.5677101	-4.36446306	-0.00303066	-2.1 4445519
	spceclies	15.07142 13	-0.12745265	-0.63279867	1.39969937
	TV news	23.3430073	-13.6323543	-1.83665774	-0.53357 161
	applied science	29.0171926	-5.33478016	-0.35330197	4.47519478
written:total		24.3074677	-10.7341 919	-1.26828537	0.82674472
	i'rspoiises	-10.3825078	6.35289683	2.88881601	-2.46083851
	lectures	-2.91 097664	7.67925377	0.2261068	-0.65477574
	coui'ses. preseiitations	-9.953 10282	6.32303802	-0.62650913	4.74424374
	talks	-10,5615833	5,0251022	-0.84427013	-0.673941 13
	news	10.9582372	-3.34438549	0.37030004	1.7984393
	sermons	-6.1 757523	0.685591 79	7,55585622	-1.67005743
	speeclies	3.17090854	1.93386535	0.85328597	0,4910585 1
	interviews	-15.0671657	-1.05027663	-1.11885311	-1.52516046
	meetings. debates	-0.3 1334219	-2.84701657	-0.46482087	0,43387244
	court/parliament cases	0.665852 14	6.0499 1021	0,98774435	-0.921 97894
spoken:total		-6,92480185	3,05799654	0,36131385	-0,23552611