JAVIER PÉREZ GUERRA
Universidad de Vigo
Vigo, España
jperez@uvigo.es

La metodología multidimensional al servicio de la determinación del grado de 'oralidad' de un tipo de texto

1. Introducción[1]

En este estudio cuestionamos la importancia de aquellas variables que tradicionalmente han sido empleadas en la caracterización del grado de oralidad de un texto (o tipo de texto) mediante la aplicación de una metodología multifactorial multidimensional. En un corpus de aproximadamente 4,5 millones de palabras, con transcripciones de inglés escrito para ser leído, inglés escrito para ser pronunciado e inglés oral, producidas por hablantes nativos británicos, investigamos la frecuencia de variables lingüísticas con el objetivo de comprobar su fiabilidad como indicadores (positivos o negativos) de la oralidad de un texto.

La metodología empleada está basada en el modelo de Biber (1988, 1995), cuya asunción fundamental es la posibilidad de caracterizar un tipo de texto mediante escalas estadísticas creadas en torno a las frecuencias de factores (o variables) lingüísticos[2] que bien tienden a co-aparecer o bien son excluyentes en un mismo grupo de textos, así como mediante los resultados del análisis multifactorial asociado a los valores concretos de dichas frecuencias. Los conjuntos de factores que conducen a resultados similares están en este modelo asociados con una interpretación funcional o 'dimensional', del tipo 'carácter informativo versus estilo dinámico implícito', 'precisión sintáctica versus léxica', 'estilo factual versus hipotético' o 'estilo planificado versus interacción en línea'.

2. El corpus

El material textual ha sido tomado de *The British National Corpus*, una colección de textos anotados del inglés contemporáneo. Dicha anotación ha permitido la extracción automática de muchos de los rasgos lingüísticos que han sido analizados en este trabajo. Hemos confiado en las etiquetas 'spolog1', 'wrimed5' y 'wriaty3/wridom/wrimed' para la clasificación de los textos en géneros, registros o tipos de textos[3], respectivamente, monólogos hablados, material escrito-para-ser-hablado y textos escritos-para-ser-leídos. La Tabla 1 ofrece los subtotales de palabras en cada clase de textos, los cuales garantizan representatividad estadística, a la luz de trabajos previos que emplean la misma metodología (desde 135.000 palabras en Kim y Biber 1994 a 960.000 palabras en Biber 1988).

Tabla 1: El corpus

categorías textuales	taxonomías		nº de palabras
hablado	respuestas clases magistrales cursos, presentaciones charlas noticias sermones	discursos entrevistas reuniones debates parlamento, corte	
subtotal			1.519.884
escrito-para-ser-hablado	oraciones sermones discursos	noticias de televisión ciencias aplicadas	
subtotal			1.370.870
escrito-para-ser-leído	imaginativo ciencias naturales ciencias aplicadas ciencias sociales actualidad internacional	comercio arte pensamiento ocio	
subtotal			1.601.832
Total			4.492.586

[1] Esta investigación está financiada por el Ministerio de Educación y Ciencia de España (proyecto HUM2005-02351/FILO), de cuyo agradecimiento deseamos dejar constancia, así como de la colaboración del profesor Xavier Gómez Guinovart (Departamento de Traducción y Lingüística de la Universidad de Vigo) en tareas de estilométrica y estadística. Pérez-Guerra (2006) constituye una versión ampliada de este trabajo en lengua inglesa.

[2] En palabras de Biber (1989: 5), "[f]unctionally based typologies have not been successful [...] in identifying the salient LINGUISTIC differences among texts in English", de ahí la necesidad de una metodología estrictamente basada en evidencia lingüística de naturaleza empírica.

[3] En este trabajo emplearemos los términos 'registro', 'género' y 'tipo de texto' como sinónimos. Para definiciones más precisas de dichos conceptos, véanse, entre otros, Wikberg (1992) y Biber (1994: 51-53).

3. Los rasgos lingüísticos

En esta investigación emplearemos rasgos lingüísticos que pueden ser contabilizados empíricamente de un modo universal, esto es, ragos cuya productividad estadística no depende de consideraciones o abstracciones teóricas. Para su identificación, hemos hecho uso, además de estudios de Biber ya mencionados, de trabajos de estilométrica y autoría (Fowler y Fowler 1922, Morton 1986, McMenamin 1993, Holmes 1994), así como otros relativos a la caracterización lingüística de textos orales y escritos (Tannen 1982, Halliday 1989, Coulthard 1985/1992, Taylor y Cameron 1987, Atkinson y Biber 1994, Hughes 1996, Eggins y Martin 1997).

Mencionamos a continuación los ragos considerados en este estudio (en muchas ocasiones hemos optado por mantener la terminología inglesa):

(i) especificidad léxica:
- ratio 'type-token': proporción de palabras diferentes
- 'hapax legomena': palabras que aparecen en una sola ocasión
- 'hapax dislegomena': palabras que aparecen en sólo dos ocasiones
- caracteres por palabra (longitud)

(ii) clases léxicas:
- 'downtoners': *almost*, *barely*, *only*, etc.
- amplificadores: *absolutely*, *completely*, *greatly*, etc.
- partículas discursivas: *well*, *now*, *anyway*, *ok*, etc.
- conjuntos: *consequently*, *furthermore*, *moreover*, *for example*, etc.
- 'hedges': *more or less*, *almost*, *maybe*, *sort/kind of*, etc.
- 'fumbles': *oh*, *er*(*m*), (*h*)*mm*, *I mean*, *actually*, *in fact*, *you know*, etc.
- 'committers': *I think*, *obviously*, *really*, etc.
- 'gambits': *in my opinion*, *to be realistic*, etc.
- 'overstaters': *purely*, *terribly*, etc.
- frases conclusivas: *something like this*, *this sort of thing* etc.
- interjecciones

(iii) adverbiales de lugar y tiempo

(iv) coordinación:
- coordinación frasal: nombre & nombre, adjectivo & adjectivo, etc.
- coordinación oracional

(v) subordinación:
- concesiva: *although*, *though*
- causativa: *because*
- condicional: *if*
- final: cláusulas con infinitivo con *to*
- cláusulas-*that* de relativo
- cláusulas-*wh* de relativo
- cláusulas de relativo del tipo 'pied-piping': *the man at whom I was looking*

(vi): formas verbales de tiempo y aspecto

(vii) modales

(viii) negación:
- analítica
- sintética: *no*, *neither*, *nor*

(ix) formas reducidas y estructuras menores:
- contracciones: *'d*, *'ll*, *'m*, *n't*, etc.
- preposiciones finales en contextos de 'preposition stranding': *the man I was looking at*
- infinitivos separados: *She has tried to consciously stop worrying*

(x) pronombres

(xi) preposiciones, adjetivos y adverbios

(xii) nombres y rasgos nominales:
- nombres abstractos, terminados en *-tion*, *-ment*, *-ness*, etc.
- nombres propios
- nombres con caso genitivo
- ratio de artículos definidos por artículos indefinidos

(xiii) preguntas directas

(xiv) pasivas:

(xv) existenciales con *there*

(xvi) longitud de la oración:
- por número de palabras
- por cláusulas

4. El proceso estadístico

La técnica estadística a la que se someten los datos crudos constituye el aspecto central de la metodología multidimensional. El objetivo de dicha técnica es el agrupamiento de rasgos en conjuntos, según su frecuencia positiva o negativa en los textos analizados. La interpretación de los grupos nos conducirá a la caracterización situacional y lingüística de cada registro o tipo de texto (Biber 1994: 32).

Resumimos a continuación los estadios principales del proceso estadístico:

(i) totales de aparición de los rasgos, conseguidos en muchas ocasiones de modo automático mediante aplicaciones AWK

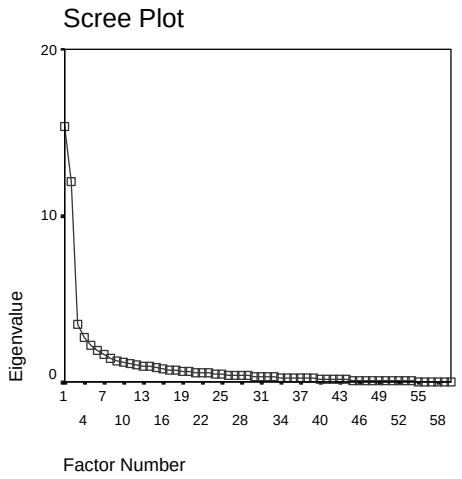
(ii) normalización estadística de la longitud de cada texto a 1.000 palabras

(iii) análisis factorial, en el que hemos usado Principal Axis Factoring para la extracción y 3-Kappa Promax para la rotación, con el fin de, por un lado, eliminar aquellos rasgos lingüísticos cuya aparición no es significativa en un factor y, por otro, de determinar cuáles son los factores que caracterizan la mayor parte de la variación. Como se puede observar en la Tabla 2 y en el Gráfico 1, de los 12 posibles factores reconocidos por el análisis factorial, 4 factores cubren más del 50% del porcentaje acumulado de variación.

Tabla 2: Variación acumulada

Factor	% de variación	% acumulado	Factor	% de variación	% acumulado
1	25,242	25,242	7	2,159	62,313
2	19,761	45,003	8	1,797	64,110
3	5,246	50,249	9	1,461	65,571
4	4,164	54,413	10	1,147	66,717
5	3,287	57,700	11	1,109	67,827
6	2,454	60,154	12	0,890	68,716

Gráfico 1: Variación acumulada –‘scree plot’–



(iv) selección de los rasgos con cargas factoriales absolutas superiores a 0,35, tal y como se defiende en Biber (1988)

(v) ordenamiento de los rasgos según sus cargas factoriales absolutas y su polaridad (positiva cuando la aparición del rasgos es frecuente en el factor; negativa cuando la frecuencia del rasgo no es relevante para la caracterización del factor), tal y como se muestra en la Tabla 3:

Tabla 3: Grupos de rasgos lingüísticos o ‘factores’ (carga factorial > 0,35)

Factor 1				Factor 2			
adjetivos atributivos	+	(existenciales)	-	elementos-*wh*	+	(‘type/token’)	-
pasivas con agente-*by*	+	(‘hedges’)	-	adverbios	+	(‘hapax legomena’)	-
caracteres/palabra	+	(modales predicativos)	-	demostrativos	+	(palabras/clausula)	-
nombres	+	frases conclusivas	-	Adjetivos predicativos	+	caracteres/palabra	-
sintagmas preposicionales	+	tiempo presente	-	sintagmas preposicionales	+		
pasivas sin agente	+	condicionales	-	modales de posibilidad	+		
palabras/cláusula	+	causativos	-	pronombre *it*	+		
coordinación frasal	+	aspecto progresivo	-	‘pied-piping’	+		
abstractos	+	pronombres indefinidos	-	Cláusulas con *to*	+		
nombres propios	+	preguntas directas	-	coordinación frasal	+		
genitivos	+	interjecciones	-	Pasado	+		
‘hapax legomena’	+	‘committers’	-	conjuntos	+		
conjuntos	+	negación analítica	-	Nombres	+		
‘type/token’	+	pronombre *it*	-	pasivas sin agente	+		
‘downtoners’	+	partículas discursivas	-	pronombres indefinidos	+		
‘pied-piping’	+	‘fumbles’	-	Negación analítica	+		
concesivas	+	pronombres de 1ª pers.	-	amplificadores	+		
adverbios de lugar	+	contracciones	-	pronombres de 3ª pers.	+		
adjetivos predicativos	+	pronombres de 2ª pers.	-	Adjetivos atributivos	+		
				contracciones	+		
				Presente	+		
				relativo *that*	+		
				Perfecto	+		
				existenciales	+		
				pronombres de 2ª pers.	+		
				abstractos	+		
				(pasivas con agente-*by*)	+		
				(‘committers’)	+		
				(adverbios de tiempo)	+		
				(‘downtoners’)	+		
				(‘fumbles’)	+		
				(partículas discursivas)	+		
				(condicionales)	+		

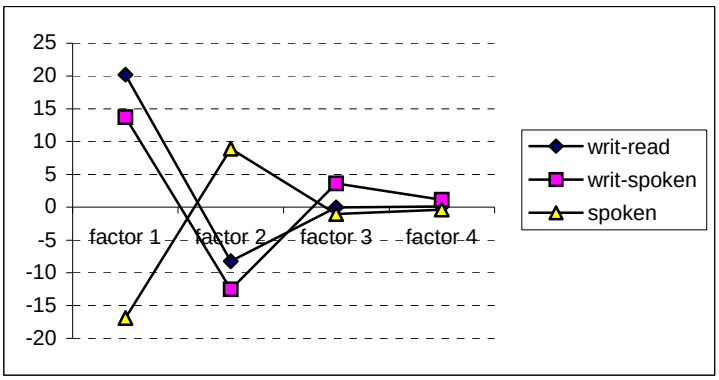
Factor 3				Factor 4			
pasado	+	(modales de necesidad)	-	cláusulas/oración	+	(interjecciones)	-
pronombres de 3ª pers.	+	(condicionales)	-	palabras/cláusula	+	(‘fumbles’)	-
nombres propios	+	(modales de posibilidad)	-	(progresivo)	+		
adverbios de lugar	+	(abstractos)	-				
(adverbios de tiempo)	+	relativo *that*	-				

(vi) obtención de los valores estandarizados de cada rasgo (estandarización a una media de 0,0 y una desviación estándar de 1,0)

(vii) obtención del valor de factor (suma de los valores estandarizados positivos y resta de los negativos)

(viii) obtención del valor del género (media de los valores de factor de todos los textos del mismo género), tal y como se muestra en el Gráfico 2 (‘writ-read’ = escrito-para-ser-leído, ‘writ-spoken’ = escrito-para-ser-hablado, ‘spoken’ = hablado):

Gráfico 2: Géneros ponderados por factores



5. Interpretación dimensional y conclusiones

Biber (1994: 35) mantiene que "no single dimension is adequate in itself to account for the range of linguistic variation in a language: rather, a multidimensional analysis is required". La técnica empleada en este trabajo ha ponderado tres tipos de textos según tres factores o 'dimensiones' resultantes de un agrupamiento de rasgos lingüísticos. El estudio pormenorizado de los rasgos –tanto positivos como negativos– de cada factor conduce a la interpretación funcional o dimensional de dichos factores:

- factor o dimensión 1: 'estilo informativo versus dinámico implícito'
- factor o dimensión 2: 'precisión sintáctica versus léxica'
- factor o dimensión 3: 'estilo factual versus hipotético'
- factor o dimensión 4: 'estilo planificado versus interacción en tiempo real ('on-line')'

El Gráfico 2 superior representa el comportamiento de las tres categorías textuales (escrito-para-ser-leído, escrito-para-ser-hablado y hablado) en relación con los cuatro factores o dimensiones resultantes del análisis multidimensional. Por lo que respecta a la dimensión 1, las dos categorías de textos escritos (tanto para-ser-leídos como para-ser-hablados) tienen valores positivos altos, asociados con estilo informativo, mientras que el estilo de los textos hablados queda caracterizado como dinámico e implícito. En cuanto a la dimensión 2, los textos escritos presentan estrategias lingüísticas asociadas con la precisión léxica. Dicha precisión posee carácter sintáctico en los textos hablados. Finalmente, las dimensiones 3 y 4, con cargas factoriales más bajas, son, estadísticamente hablando, menos relevantes que las dimensiones ya referidas. Es por ello por lo que el comportamiento de las tres categorías textuales es muy similar en las dimensiones 3 y 4.

A modo de conclusión, las variables funcionales más significativas en la caracterización de los textos escritos versus orales son aquellas que están asociadas con las dimensiones 1 y 2, esto es, estilo informativo y precisión léxica en la lengua escrita, y estilo dinámico implícito y precisión sintáctica en el material hablado.

Referencias

Aston, Guy y Lou Burnard 1997. *The BNC handbook. Exploring the British National Corpus with SARA*. Edimburgo: Edinburgh University Press.

Atkinson, Dwight y Douglas Biber 1994. Register: a review of empirical research. Eds Biber y Finegan, págs. 351-385.

Biber, Douglas 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber. Douglas 1989. A typology of English texts. *Linguistics* 27/1, págs. 3-43.

Biber, Douglas 1994. An analythical framework for register studies. Eds Biber y Finegan, págs. 31-56.

Biber, Douglas 1995. *Dimensions of register variation. A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas y Edward Finegan eds 1994. *Sociolinguistic perspectives on register*. Nueva York: Oxford University Press.

Coulthard, Malcolm 1985 [1977]. *An introduction to discourse analysis*. Londres: Longman.

Coulthard, Malcolm ed 1992. *Advances in spoken discourse analysis*. Londres: Routledge.

Eggins, Suzanne y J.R. Martin 1997. Genres and registers of discourse. Ed Teun A. Van Dijk. *Discourse as structure and process. A multidisciplinary introduction. Volume I*. Londres: SAGE, págs. 230-256.

Fowler, Henry W. y Frank G. Fowler 1922. *The king's English*. Oxford: Clarendon.

Halliday, M.A.K. 1989 [1985]. *Speech and written language*. Oxford: Oxford University Press.

Holmes, David I. 1994. Authorship attribution. *Computers and the Humanities* 28, págs. 87-106.

Hughes, Rebecca 1996. *English in speech and writing. Investigating language and literature*. London: Routledge.

Kim, Yong-Jin y Douglas Biber 1994. A corpus-based analysis of register variation in Korean. Eds Biber y Finegan, págs. 157-181.

McMenamin, Gerald R. 1993. *Forensic stylistics*. Amsterdam: Elsevier.

Morton, A.Q. 1986. Once. A test of authorship based on words which are not repeated in the sample. *Literary and Linguistic Computing* 1/1, págs. 1-8.

Pérez-Guerra, Javier 2006. How oral is this text? Revisiting orality through multidimensional techniques. Eds Ana María Hornero, María José Luzón y Silvia Murillo. *Corpus linguistics. Applications for the study of English*. Bern: Peter Lang, págs. 483-506.

Tannen, Deborah 1982. Oral and literate strategies in spoken and written narratives. *Language* 58/1, págs. 1-21.

Taylor, Talbot J. y Deborah Cameron 1987. *Analyzing conversation. Rules and units in the structure of talk*. Oxford: Pergamon Press.

Wikberg, Kay 1992. Discourse category and text type classification: procedural discourse in the Brown and the LOB corpora. Ed Gerhard Leitner. *New directions in English language. Methodology, results, software developments*. Berlín: Mouton, págs. 247-261.