INTERDISCIPLINARIEDAD LENGUAS Y TIC: INVESTIGACIÓN Y ENSEÑANZA

IULMA Institut Interuniversitari de Llengües Modernes Aplicades de la Comunitat Valenciana
Instituto Interuniversitario de Lenguas Modernas Aplicadas de la Comunidad Valenciana

UNIVERSITAT JAUME I · Universitat d'Alacant Universidad de Alicante · VNIVERSITAT DE VALÈNCIA

Universitat de València, 10-12 marzo de 2010

Measurements in the degree of specialisation of a text: how to interpret the Type Token Ratio

Nuria Edo Marzá
Universitat de València – IULMA
nuria.edo@uv.es

Resumen

El presente artículo tiene como objetivo proporcionar claves a la hora de medir e interpretar el grado de especialización de una muestra textual tomando como base un valor numérico, el Ratio Tipo Item (RTI). Este valor, obtenido en esta investigación con el software de concordancias WordSmith Tools 4.0, presenta, sin embargo, una serie de particularidades que hacen necesario un "manejo cuidadoso" del mismo, siendo necesario el ser plenamente consciente de las limitaciones de la información que proporciona.

Esto es así porque el RTI, centro de este estudio y, en menor medida, su equivalente más equilibrado, el Ratio Tipo Item Estandarizado (RTIE), son medidas inestables ampliamente dependientes del tamaño de las muestras textuales (medidas en cuanto a "ítems" (tokens) o palabras totales en el texto). En consecuencia, para fines comparativos, estos ratios, especialmente el RTI, dependen fuertemente del tamaño de las muestras textuales analizadas, siendo fiables sólo en aquellos casos en que estas muestras tienen un tamaño similar. Así pues, este artículo tiene como objetivo proporcionar algunas claves para interpretar de manera plausible estos ratios, tomando siempre en consideración factores como el tamaño, que, en caso de no ser contemplados, pueden hacer la comparación poco fiable e inútil. Por lo tanto, pese a que el RTI no constituye una verdad absoluta en sí mismo, puede proporcionar información valiosa para el terminógrafo si se interpreta de manera correcta.

La interpretación empírica proporcionada aquí para este valor numérico puede ser reveladora debido al importante rol que el RTI puede jugar en la compilación y desarrollo de córpora equilibrados, en el estudio de género (y del nivel de especialización esperable en cada género) y en el análisis de los lenguajes especializados, entre otras aplicaciones. De ello, y con la base de un corpus en cerámica industrial, nos hemos centrado en el estudio de cómo usar e interpretar el RTI para ayudarnos a alcanzar un equilibrio textual en el corpus.

Esto puede hacerse mediante la compilación de muestras textuales con diferentes grados de espcialización que contribuyan a lograr un equilibrio textual representativo. Este equilibrio, basado en la inclusión de textos con distintos grados de especialización (y diferentes géneros), refleja mejor los dominios de especialidad y subyace en la representatividad del corpus, tan necesaria en estudios de este tipo. Como resultado, esta investigación plantea una interpretación empírica del RTI en forma de intervalos de tamaño – establecidos sobre la base de las muestras analizadas en este estudio – para las cuales 3 valores distintos ("altamente especializado", "especializado" y "moderadamente especializado") han sido asignados según el caso.

Palabras clave: Ratio Tipo Item (RTI), grado de especialización, corpus equilibrado, representatividad, Ratio Tipo Item estandarizado (RTIE), equilibrio.

Abstract
This paper is intended at providing hints for measuring and interpreting the degree of specialisation of a textual sample on the basis of a numerical value, the Type Token Ratio (TTR). This value, obtained in this research with the concordance software program WordSmith Tools 4.0, presents, however, a series of shortcomings which makes it necessary to "handle it with care", being aware of the limitations of the information it provides. This is so since the TTR, focus of this study, and, in a lesser degree, its more balanced counterpart, the Standardised Type Token Ratio (STTR), are unstable measurements deeply dependent on the size of the textual samples (measured in terms of "tokens" or running words in the text). Consequently, for comparison purposes, these ratios, specially the TTR, are deeply dependent on the size of the textual samples analysed, being reliable only in those cases in which these samples are of a similar size. Accordingly, this article aims at providing some hints for a plausible numerical interpretation of these ratios, bearing always in mind factors such as size which, if disregarded, can make comparison unreliable and useless. Therefore, even when the TTR does not constitute an absolute truth in itself, it can provide valuable information for the terminographer if correctly and cautiosly interpreted.

The empirical interpretation provided here to this numerical value may be enlightening due to the active and important role the TTR may play in the compilation and development of balanced corpora, in the study of genre (and the expectable degree of specialisation of each genre) and in the analysis of specialised languages among other applications. From this, and on the basis of a corpus on industrial ceramics, we have focused on the study of how to use and interpret the TTR in order to help us achieve textual equilibrium in a corpus. This can be done by means of compiling textual samples with different degrees of specialisation leading to the achievement of a representative textual balance. This balance based on the inclusion of texts showing different degrees of specialisation (and different genres) reflects speciality domains better and accounts for the features of representativeness and equilibrium so necessary in corpus studies. As a result, this study poses an empirical interpretation to the TTR in the form of a series of size intervals – established on the basis of the samples analysed in this study – to which 3 different numerical values representing the specialisation degrees of "highly specilised", "specialised" and "fairly specialised" have been assigned.

Key words: Type Token Ratio (TTR), degree of specialisation, balanced corpus, representativeness, Standardised Type Token Ratio (STTR), equilibrium.

Referencias bibliográficas / Bibliographical references:

Barber, C. L. (1962). "Some measurable characteristics of modern English scientific Prose". *Gothenburg Studies in English* 14 (pp.21-43).

Cabré Castellví, M.T. (1993). *La Terminología : Teoría, metodología, aplicacione.,* Editorial Antártida/Empúries.

Cabré Castellví, M. T. (1999). *La terminología, representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada (IULA), Universitat Pompeu Fabra.

Cabré Castellví, M.T. (2002). "Textos especializados y unidades de conocimiento: metodología y tipologización", in J. García Palacios and Fuentes Morán, M. T. (Eds.): *Texto, terminología y traducción*, Salamanca: Almar.

Ciapuscio, G.E. and Kugel, I. (2002). "Hacia una tipología del discurso especializado: aspectos teóricos y aplicados", in García Palacios, J. and Fuentes Morán, M.T. (Eds.): *Texto, terminología y traducción* (pp. 37-73). Biblioteca de Traducción, Salamanca: Ediciones Almar, 37-73.

Edo Marzá, N. (2008). "The Communicative Theory of Terminology (CTT) applied to the development of a corpus-based specialised dictionary of the ceramics industry". Unpublished PhD dissertation, Castelló, Spain.

Granger, S. and Wynne, M. (2000). "Optimising Measures of Lexical Variation in EFL Learner corpora". In J. M. Kirk. (Ed.) Corpora Galore. Amsterdam: Rodopi (pp. 249-258). Online document: http://www.fltr.ucl.ac.be/fltr/germ/etan/CECL/Downloads/GrangerWynnerodopi998.doc [Last search date: November 2008].

Hoffmann, L. (1987). “Language for Special/Specific Purposes”, in Ammon, U., Ditmar, N. and Matheier, K. J. (Eds.): *Sociolinguistics. An International Handbook of the Science of Language and Society*, Berlin/New York: de Gruyter, (pp. 298-302).

IULA (2007). “Terminología y enseñanza de lenguas”, in Grup IulaTerm*, Diploma de postgrado online: Terminología y necesidades profesionales,* Barcelona: IULA, Universidad Pompeu Fabra. Online document: http://www.iulaonline.iula.upf.edu, ISBN 84-89782-01-6. [Last search date: October 2008].

Pérez Hernández, M. Ch. (2002). “Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento”, *Estudios de Lingüística Española* (*ELiEs*), 18. Online document: http://elies.rediris.es/elies18/index.html [Last search date: December, 2008].

Sager J.C., D. Dungworth, P.F. McDonald (1980). *English special languages. Principles and practice in science and technology*, Wiesbaden, Brandstetter.

Scott, M. and Oxford University Press (1998). *WordSmith Tools Manual version 4.0*. Online document: http://www.lexically.net/wordsmith/version4/manual.pdf [Last search date: November 2008].

Vargas Sierra, Ch. (2005). *Aproximación terminográfica al lenguaje de la piedra natural. Propuesta de sistematización para la elaboración de un diccionario traductológico*, PhD dissertation, Alicante, Spain.

Vargas Sierra, C. (2006). “Diseño de un corpus especializado con fines terminográficos: el corpus de la Piedra Natural”, *Debate Terminológico, 2* (7/2006). París: RITERM (Red Iberoamericana de Terminología).

Wüster, E. (1979): *Einführung in die allgemanie Terminologielehre und terminologische Lexikographie*, Wien: Springer (Spanish version by M. T. Cabré Castellví: *Introducción a la teoría de la terminología y a la lexicografía terminológica.* Barcelona: IULA, Universitat Pompeu Fabra, 1998).