Listas de frequência de palavras como marcadores de estilo no reconhecimento de autoria

Rui Sousa Silva
Universidade do Porto
rmsilva@me.com

O estudo e análise do discurso tem sido objecto de diferentes teorias e abordagens (Coulthard, 1977; Dijk, 1997; Fairclough & Wodak, 1997; Sinclair, 1991) desde a análise da interacção entre o discurso e a sociedade (Dijk, 1997; Fairclough & Wodak, 1997) à análise do discurso enquanto realização linguística (Coulthard, 1977; Sinclair, 1991), passando pelo estudo da relação entre a linguística e a lei como forma de linguística forense (análise forense do discurso) (Coulthard & Johnson, 2007). É neste contexto que se insere o reconhecimento de autoria.

Neste estudo procuramos verificar a utilidade e aplicabilidade das listas de frequências de palavras como marcador de discurso no reconhecimento de autoria em português, a exemplo do que acontece com outras línguas (Hänlein, 1998), utilizando conceitos de estilística forense (McMenamin, 2002). Neste sentido, recorremos aos estudos em linguística com corpos (Biber, Conrad, Reppen, & Aitchison, 2000; McEnery & Wilson, 2001) para criar um corpo de textos com cerca de 100.000 palavras, escritos por dois cronistas diferentes (António Barreto e José Pacheco Pereira), publicados no jornal *Público* entre Janeiro e Dezembro de 2007. Utilizando o Corpógrafo (Maia, Sarmento & Santos, 2005), analisamos a frequência de palavras, a frequência de utilização de palavras utilizadas pelo autor uma única vez (*hapax legomena*) e a frequência das palavras que ocorrem duas vezes no texto do mesmo autor (*hapax dislegomena*). Para verificar os resultados do presente estudo, analisamos dois textos escritos pelos dois autores, publicados no jornal *Público* em 2008, confrontando, assim, os textos com as conclusões do estudo do corpo de textos.

Em conclusão, esta análise deverá demonstrar a utilidade das listas de frequência de palavras como critério de reconhecimento de autoria em português. Os resultados do estudo permitirão verificar que, uma vez que cada autor possui um *idiolecto* próprio (Coulthard & Johnson, 2007), com marcas de autoria distintas, diferentes textos, produzidos por diferentes autores, recorrem à utilização de elementos idiossincráticos e padrões linguísticos distintos.

Bibliografia

Biber, Douglas, Conrad, Susan & Reppen, Randi. (2000). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Coulthard, Malcolm. (1977). *An Introduction to Discourse Analysis*. Londres: Longman.

Coulthard, Malcolm, & Johnson, Alison. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. Londres e Nova Iorque: Routledge.

Dijk, Teun A. van (1997). Discourse as Interaction in Society. In Teun A. van Dijk (Ed.), *Discourse Studies: A Multidisciplinary Introduction - Discourse as Social Interaction* (Vol. 2, pp. 1-37). Londres: SAGE Publications Ltd.

Fairclough, Norman, & Wodak, Ruth. (1997). Critical Discourse Analysis. In Teun A. van Dijk (Ed.), *Discourse Studies: A Multidisciplinary Introduction - Discourse as Social Interaction* (Vol. 2, pp. 258-284). Londres: SAGE Publications Ltd.

Hänlein, Heike. (1998). *Studies in Authorship Recognition - A Corpus-based Approach.* Francoforte: Peter Lang.

McEnery, Tony, & Wilson, Andrew. (2001). *Corpus Linguistics: An Introduction*. Edinburgo: Edinburgh University Press.

McMenamin, Gerald R. (2002). *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton e Nova Iorque: CRC Press.

Maia, Belinda, Sarmento, Luís, & Santos, Diana. (2005). Introduzindo o Corpógrafo - um conjunto de ferramentas para criar corpora especializados e comparáveis e bases de dados teminológicas. *Terminómetro - Número especial nº 7 - A terminologia em Portugal e nos países de língua portuguesa em África* (2005), pp. 61-62.

Sinclair, John M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.