Text mining e análise de géneros: contributo para a análise automática de textos na perspetiva do ISD

GONÇALVES, Matilde

NOVA FCSH/CLUNL

MAGALHÃES, Miguel

FCT/CLUNL

(Eixo 1: Teorias e epistemologias)

O termo "género" tem uma longa tradição em muitas disciplinas académicas, como os estudos literários, a linguística e a retórica. É, consequentemente, um conceito difícil de manusear, uma vez que cada uma dessas áreas de estudo tem a sua própria definição e abordagem de análise. Mais recentemente, estendeu-se a outras áreas, como *big data* recolha de informação, onde surgiu a necessidade de recolher e tratar automaticamente grandes volumes de informação (neste caso, texto). Na última década, o *text mining* teve um rápido desenvolvimento. Considerado um epifenómeno do *data mining*, um dos desafios do *text mining* é a identificação do género a que pertence o texto em análise. A diversidade de conceitos e definições do género não tem afetado este campo de pesquisa, uma vez que adotam ou adaptam definições já estabelecidas. O risco que esta adaptação tem revelado, é que não existe um consenso sobre o objeto de estudo, afetando a abordagem prática dos textos.

Nos estudos linguísticos, a questão do género tem desempenhado um papel central e, atualmente, podem ser identificadas várias linhas teóricas que abordam a questão do género de acordo com áreas ou problemas específicos: Biber (1995), Bronckart (1997, 2008), Maingueneau (2005), Rastier (1989, 2009), Swales (1990).

Face ao exposto, o primeiro objetivo desta apresentação é analisar e apresentar uma nova abordagem para a classificação automática de géneros baseada numa análise multidimensional, que a seguir se apresenta sucintamente. Partindo de uma análise da organização semântica e discursiva, e da estrutura temática de uma amostra de textos de um *corpus* escrito, vamos tentar perceber se se o comentário é uma prática textual onde coexistem diferentes géneros, ou um conjunto de textos sem fronteiras definidas. A partir desta análise, o segundo objetivo deste trabalho será o desenvolvimento de uma metodologia que permita analisar, com o auxílio de ferramentas de *text mining*, e de forma semiautomática o género dos textos.

Este trabalho parte da proposta de Santini (2005, 2007) propondo um novo conjunto de traços linguísticos com base no referencial teórico do Interacionismo Sócio-Discursivo, desenvolvido por Bronckart (1997). E mais especificamente, este trabalho analisa como os tipos discursivos, de acordo com a formulação proposta por Bronckart (2008), entendidos como materializações linguísticas dos mundos discursivos, nos permitem descrever os textos. Embora Biber (1995) já tenha proposto uma abordagem semelhante para a análise de textos orais e escritos, a nossa proposta permite uma análise mais ampla, identificando

não só os meios de comunicação, mas também a função comunicativa do texto, combinando análise qualitativa e dados quantitativos. Para isso, desenvolveu-se uma metodologia de etiquetagem de *corpus* que permite descrever os marcadores de género (Miranda, 2010) presentes no texto e introduzi-los como variáveis de análise. Deste modo, é possível materializar e quantificar as representações que o sujeito tem do contexto da ação e que são convocadas pelo agente.

Os resultados preliminares deste processo de etiquetagem mostram que uma análise baseada no quadro do ISD permite não só analisar o texto enquanto unidade comunicativa como também dá resposta à necessidade de flexibilizar a noção de género e permite integrar a heterogeneidade que os textos apresentam.

Referências:

Biber, D. (1995). *Variation across speech and writing*. Cambridge University Press.

Bronckart, J.-P. (1997). *Activité langagière, textes et discours. Pour un interactionisme socio-discursif*. Paris: Delachaux et Niestlé.

Bronckart, J.-P. (2008). Genres de textes, types de discours et "degrés" de langue : hommage à François Rastier. *Texto !*, *XIII*(1/2).

Maingueneau, D. (2005). As categorias da análise do discurso. In F. Menendez (Ed.), *Análise do discurso* (pp. 81–105). Lisboa: Huguin.

Miranda, F. (2010). *Textos e géneros em diálogo. Uma abordagem linguística da intertextualização*. Lisboa: FCG/FCT

Rastier, F. (1989). *Sens et textualité*. Paris: Hachette.

Rastier, F. (2009). Stylistique et textométrie Sept questions de principe et d'opportunité. *Texto!*, *XV*(4), 62–70.

Santini, M. (2005). Linguistic facets for genre and text type identification: A description of linguistically-motivated features. *ITRI Report Series: ITRI-05*, 1–41.

Santini, M. (2007). Automatic Genre Identification : Towards a Flexible Classification Scheme, (1), 1–6.

Swales, J. M. (1990). *Genre analysis*. Cambridge: Cambridge University Press.