Capítulo 13

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

Rui Sousa Silva

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 13, p. 93–97, 2008

O estudo e análise do discurso tem sido objecto de diferentes teorias e abordagens (Coulthard, 1977; Dijk, 1997; Fairclough e Wodak, 1997; Sinclair, 1991), desde a análise da interacção entre o discurso e a sociedade e a análise crítica do discurso (Dijk, 1997; Fairclough e 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 e Johnson, 2007; Shuy, 2006).

A análise forense do discurso, enquanto ramo da linguística aplicada, possui aplicações diversificadas, entre as quais: a identificação de autoria; a identificação do modo (no sentido de Halliday); a tradução e a interpretação jurídica; a transcrição de declarações e depoimentos; o estudo da linguagem e discurso dos tribunais; o estudo de direitos linguísticos; a análise de declarações; a fonética forense; e o estudo do estatuto textual. Neste artigo, debruçamo-nos sobre a primeira: a identificação de autoria. Recorrendo à análise da utilização da linguagem pelo autor e das informações que essa análise transmite ao analista acerca do escritor, linguisticamente (Olsson, 2004), procuramos determinar o perfil de autoria textual, isto é, identificar o autor com base numa análise contrastiva de um corpo de textos limitado (Coulthard e Johnson, 2007; Olsson, 2004).

Para determinar este perfil, não podemos limitar-nos à utilização de dados puramente estatísticos dos próprios textos estudados, uma vez que o contexto sociocultural e a realidade extra-textual influenciam a forma de falar e de escrever dos falantes de uma determinada língua; num mesmo país ou cultura, diferentes pessoas, com acesso diferente a educação, formação e informação, têm formas semelhantes de produção textual. O sociolecto (i.e., a variedade de uma língua característica de uma determinada classe ou estatuto social) pode restringir a gama possível de autores, mas não é um factor decisivo. A análise estatística dos dados constitui, assim, um dos métodos utilizados, mas não o único. Daí o recurso à análise forense do discurso como forma de equacionar os dados mais relevantes do corpo de textos.

Considerando todos estes princípios, teremos, então, que procurar identificar o idiolecto de cada um dos autores, isto é, presumindo que todos os falantes nativos de uma língua possuem uma versão distinta e individual da língua que falam e escrevem, teremos que procurar no texto marcadores que apontem para a selecção individual de aspectos linguísticos genéricos (Coulthard e Johnson, 2007). Socorremo-nos, para o efeito, de três princípios da estilística forense: o princípio de que o estilo individual de cada autor é determinado pela escolha (Hänlein, 1998); o grau em que o autor tende para determinadas formas de "expor as coisas" (McEnery e Wilson, 1996); e, finalmente, o pressuposto de que é necessário identificar um conjunto agregado e único de marcadores, presentes individualmente noutros autores (McMenamin, 2002). Reconhecendo a validade e a fiabilidade de marcadores como o formato do texto, a utilização de números/símbolos, abreviaturas, pontuação, maiúsculas/minúsculas, ortografia, formação lexical, sintaxe, discurso, er-

ros e correcção, utilização da voz activa e passiva, entre outros, focmo-nos, neste estudo, nas expressões e palavras de elevada frequência, no sentido de verificar a sua utilidade e aplicabilidade como marcador de discurso no reconhecimento de autoria em português, a exemplo do que acontece para outras línguas (Hänlein, 1998)[1].

Com base nos estudos em linguística com corpos (Biber et al., 2000; McEnery e Wilson, 1996), criámos um corpo de 84 textos escritos pelos cronistas António Barreto e José Pacheco Pereira, com 107.360 átomos, publicados no jornal Público entre Janeiro e Dezembro de 2007. Recorrendo ao Corpógrafo (Sarmento et al., 2004; Maia e Matos, 2008), analisámos a frequência de expressões com um comprimento de quatro gramas (i.e., tetragramas) utilizadas pelo autor uma única vez (*hapax legomena*) e a frequência de expressões que ocorrem mais vezes nos textos do mesmo autor (*hapax dislegomena*). Depois de proceder à extracção dos tetragramas mais frequentes, procedemos à sua classificação, manualmente, segundo uma taxonomia de 15 classes, conforme proposto por Sousa Silva (2006): *especificação*, *explicação*, *exemplificação*, *comparação*, *contraste*, *generalização*, *correcção*, *preparação*, *inclusão*, *concessão*, *restrição*, *enumeração*, *propósito*, *negação*, *justificação*. Os resultados desta análise, apresentados nas tabelas 13.1 (com uma ordenação por classe semântica) e 13.2 (com uma ordenação por frequência decrescente de utilização), mostram que os dois autores recorrem a estratégias semânticas de produção textual diferentes. Os valores classificados como ruído resultam de n-gramas obtidos com caracteres não reconhecidos – e, por isso, considerados erros.

Comparando a utilização das classes pelos dois autores, verificamos, conforme apresentado na tabela 13.1, que os dois autores recorrem com uma frequência idêntica a estratégias de *correcção*, *negação* e *restrição*, utilizando, porém, de forma distinta as restantes classes:

A interpretação que fazemos dos dados obtidos permite-nos constatar que, enquanto António Barreto recorre a expressões com um valor semântico que lhe permitem ser mais claro, directo e focalizado, José Pacheco Pereira apresenta características de uma produção textual mais vaga, hesitante e inconstante -– frequentemente conotada com uma literacia elitista.

Para verificar os resultados do presente estudo, analisámos dois textos escritos pelos dois autores, publicados no jornal Público em 2008. A metodologia adoptada consiste na aplicação de um "teste cego" (isto é, com textos cuja autoria foi tornada anónima), com o objectivo de confrontar os textos com as conclusões do estudo do corpo de textos. Considerando que estes dois textos são demasiado pequenos para uma análise estatística (cerca de mil átomos por texto), procurámos traços individuais marcantes em cada um deles, nomeadamente a frequência das palavras utilizadas no corpo de textos recolhidos em 2007

[1] Neste contexto, entendemos "palavras" no sentido que lhe foi atribuído por Halliday (1994) de "wordings", ou seja, são palavras as sequências gramaticais, ou "sintagmas", constituídas por elementos de três tipos: elementos lexicais (tais como verbos e nomes), elementos gramaticais (tais como artigos e determinantes), e elementos intermédios (tais como preposições) – todos eles elementos que constituem os n-gramas.

António Barreto			José Pacheco Pereira		
Classe	Total	%	Classe	Total	%
comparação	20	5,13	comparação	50	10,22
concessão	8	2,05	concesão	13	2,66
contraste	41	10,51	contraste	17	3,48
correcção	0	0,0	correcção	0	0,0
enumeração	24	6,15	enumeração	63	12,88
exemplificação	9	2,31	exemplificação	11	2,25
explicação	18	4,62	explicação	91	18,61
generalização	18	4,62	generalização	8	1,64
inclusão	16	4,10	inclusão	4	0,82
justificação	0	0,0	justificação	10	2,04
negação	0	0,0	negação	0	0,0
preparação	10	2,56	preparação	8	1,64
propósito	8	2,05	propósito	6	1,23
restrição	0	0,0	restrição	0	0,0
especificação	218	55,90	especificação	208	42,54
Total	**390**	**100,0**	**Total**	**489**	**100,0**
ruído	**0**		**ruído**	**1**	

Tabela 13.1: Lista comparativa de classes semânticas utilizadas pelos autores (ordenadas por classe semântica).

António Barreto			José Pacheco Pereira		
Classe	Total	%	Classe	Total	%
especificação	218	55,90	especificação	208	42,54
contraste	41	10,51	explicação	91	18,61
enumeração	24	6,15	enumeração	63	12,88
comparação	20	5,13	comparação	50	10,22
explicação	18	4,62	contraste	17	3,48
generalização	18	4,62	concesão	13	2,66
inclusão	16	4,10	exemplificação	11	2,25
preparação	10	2,56	justificação	10	2,04
exemplificação	9	2,31	generalização	8	1,64
concessão	8	2,05	preparação	8	1,64
propósito	8	2,05	propósito	6	1,23
correcção	0	0,0	inclusão	4	0,82
justificação	0	0,0	correcção	0	0,0
negação	0	0,0	negação	0	0,0
restrição	0	0,0	restrição	0	0,0
Total	**390**	**100,0**	**Total**	**489**	**100,0**
ruído	**0**		**ruído**	**1**	

Tabela 13.2: Lista comparativa de classes semânticas utilizadas pelos autores (ordenadas por frequência).

António Barreto			José Pacheco Pereira	
−	comparação	≠	+	comparação
−	concessão	≠	+	concessão
+	contraste	≠	−	contraste
−	**correcção**	=	−	**correcção**
−	enumeração	≠	+	enumeração
+	exemplificação	≠	−	exemplificação
−	explicação	≠	+	explicação
+	generalização	≠	−	generalização
+	inclusão	≠	−	inclusão
−	justificação	≠	+	justificação
−	**negação**	=	−	**negação**
+	preparação	≠	−	preparação
+	propósito	≠	−	propósito
−	**restrição**	=	−	**restrição**
+	especificação	≠	−	especificação

Tabela 13.3: Comparação das classes semânticas utilizadas pelos dois autores.

(e que aqui utilizamos como corpo de referência, isto é, com o corpo de textos com o qual comparamos os textos A e B). A lista de frequência de palavras dos textos anónimos referidos como "Texto A" e "Texto B" mostra que, enquanto o Autor A utiliza com maior frequência as expressões "acima de tudo," e "o que significa que", o Autor B utiliza expressões como "a verdade é que", "ao mesmo tempo que" e "assim como o de". Contrastando estes resultados com os resultados obtidos na análise do corpo de textos utilizado no estudo, verificamos que as expressões utilizadas pelos autores A e B correspondem, respectivamente, a José Pacheco Pereira e António Barreto.

Este estudo permite, assim, comprovar que existem diferenças semânticas significativas, mesmo tratando-se de autores que escrevem com uma regularidade semelhante para um mesmo público, sob orientações editoriais idênticas. Poderemos, por isso, interpretar os dados obtidos como sendo um marcador de autoria válido e fiável em português, a exemplo do que acontece com outras línguas (como é o caso do inglês). Poderemos, por isso, constatar que, uma vez que cada autor possui um idiolecto próprio (Coulthard e Johnson, 2007), com marcas de autoria distintas, diferentes textos, produzidos por diferentes autores, recorrem à utilização de elementos idiossincráticos e padrões linguísticos distintos.

Em conclusão, esta análise demonstra a utilidade das listas de frequência de palavras como critério de reconhecimento de autoria em português.