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

Rui Sousa Silva
Faculdade de Letras da Universidade do Porto
rmsilva@me.com

Análise do Discurso

- Análise da interacção entre o discurso e a sociedade e a análise crítica do discurso (Dijk, 1997; Fairclough & Wodak, 1997)
- Análise do discurso enquanto realização linguística (Coulthard, 1977; Sinclair, 1991)
- Análise forense do discurso (relação entre a linguística e a lei como forma de linguística forense (Coulthard & Johnson, 2007)

Perfis de autoria

- Estudo consiste em analisar a utilização da linguagem pelo autor, e as informações que isso transmite ao analista acerca do escritor, linguisticamente (Olsson, 2004)
- Atribuição de autoria: resolver disputas, determinar autor de textos anónimos
- Identificação do autor: determinar o autor com base numa análise contrastiva de um corpo de textos limitado (Olsson, 2004; Coulthard & Johnson, 2007)

Estilística Forense

- Estilo individual de cada autor é determinado pela escolha (Hänlein, 1999)
- Grau em que o autor tende para determinadas formas de "pôr as coisas" (McEnerty & Wilson, 1996)
- Necessário identificar um conjunto agregado (único) de marcadores, presentes individualmente noutros autores (McMenamin, 2002)

Marcadores de discurso

- Formato do texto;
- Números/símbolos;
- Abreviaturas;
- Pontuação;
- Uso de maiúsculas/ minúsculas;
- Ortografia;
- Formação lexical;
- Sintaxe;
- Discurso;
- Erros e correcção;
- Expressões e palavras de elevada frequência

(McMenamin, 2002)

Factores Externos

- Contexto socio-cultural, realidade extra-textual e sociolecto:
  - influenciam forma de falar/escrever dos seus falantes
  - no mesmo país ou cultura, diferentes pessoas, com acesso diferente a educação e informação, têm formas semelhantes de produção textual
  - sociolecto (variedade de uma língua característica de uma determinada classe ou estatuto social) pode restringir gama possível de autores, mas não é factor decisivo

Factores Externos

- Idiolecto

  - noção de que todos os falantes nativos de uma língua possuem uma versão distinta e individual da língua que falam e escrevem – selecção individual

  (Coulthard, 2007)

Questão de partida

- As listas de frequências de palavras poderão funcionar como marcadores de estilo no reconhecimento de autoria?

Metodologia

- Palavras no sentido de "wordings" (Halliday, 1994):

  sequências gramaticais, ou "sintagmas", constituídas por elementos de dois tipos: elementos lexicais (e.g. *v* e *n*) elementos gramaticais (e.g. *art* e *det*) e elementos intermédios (e.g. *prep*) – n-gramas

Análise

- Corpus:

  corpo de textos finito
  textos: artigos de opinião
  dois autores
  publicados no jornal diário *Público*
  data: Janeiro/Dezembro de 2007

Constituição do corpo

António Barreto	José Pacheco Pereira
41.321 átomos 37 textos	66.032 átomos 47 textos
4-gramas	4-gramas

Análise do Corpo de Textos

- Estudo de n-gramas (ordenado por frequência) utilizando o Corpógrafo: 4-gramas mais utilizados
- Classificação semântica – taxonomia de 15 classes:

  especificação, explicação, exemplificação, comparação, contraste, generalização, correcção, preparação, inclusão, concessão, restrição, enumeração, propósito, negação, justificação

Resultados da Análise

António Barreto		
Classe	Total	%
comparação	20	5,13
concessão	8	2,05
contraste	41	10,51
correcção	0	0,00
enumeração	24	6,15
exemplificação	9	2,31
explicação	18	4,62
generalização	18	4,62
inclusão	16	4,10
justificação	0	0,00
negação	0	0,00
preparação	10	2,56
propósito	8	2,05
restrição	0	0,00
especificação	218	55,90
	390	**100,00**
ruído	**0**	

José Pacheco Pereira		
Classe	Total	%
comparação	50	10,22
concessão	13	2,66
contraste	17	3,48
correcção	0	0,00
enumeração	63	12,88
exemplificação	11	2,25
explicação	91	18,61
generalização	8	1,64
inclusão	4	0,82
justificação	10	2,04
negação	0	0,00
preparação	8	1,64
propósito	6	1,23
restrição	0	0,00
especificação	208	42,54
	489	**100,00**
ruído	**1**	

% ocorrências/classes

António Barreto		
Classe	Total	%
especificação	218	55,90
contraste	41	10,51
enumeração	24	6,15
comparação	20	5,13
explicação	18	4,62
generalização	18	4,62
inclusão	16	4,10
preparação	10	2,56
exemplificação	9	2,31
concessão	8	2,05
propósito	8	2,05
correcção	0	0,00
justificação	0	0,00
negação	0	0,00
restrição	0	0,00
	390	**100,00**
ruído	**0**	

José Pacheco Pereira		
Classe	Total	%
especificação	208	42,54
explicação	91	18,61
enumeração	63	12,88
comparação	50	10,22
contraste	17	3,48
concessão	13	2,66
exemplificação	11	2,25
justificação	10	2,04
generalização	8	1,64
preparação	8	1,64
propósito	6	1,23
inclusão	4	0,82
correcção	0	0,00
negação	0	0,00
restrição	0	0,00
	489	**100,00**
ruído	**1**	

Comparação de Classes

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

+ claro, directo, focalizado

+ vago, hesitante, inconstante

Validação

- Dois textos, escritos pelos mesmos autores
- Publicados no mesmo jornal (*Público*) em 2008
- Demasiado pequenos para uma análise estatística, mas cada um deles com traços individuais marcantes
- Procurar frequências de palavras utilizadas no corpo de textos de análise

Resultados da Validação

Texto Autor A

É por causa de Manuela Ferreira Leite, do PSD, de Manuel Alegre, do BE, do PCP e acima de tudo por causa de José Sócrates, e do PS entre Alberto Martins e Vitalino Canas.

30	acima de tudo ,	5	0.007

É isso que significa a credibilidade, palavra com muito mais conteúdo do que parece e que muda muito mais coisas do que se imagina

192	o que significa que	3	0.004

Resultados da Validação

Texto Autor B

Mas a verdade é que a alegada cornucópia é muito inferior ao necessário.

1	a verdade é que	14	0.033

Ao mesmo tempo que se ouvem declarações messiânicas sobre as novas fontes de energia e a poupança de combustíveis, anunciam-se mais auto-estradas, pontes e viadutos.

15	ao mesmo tempo ,	6	0.014

Prepara-se o fecho definitivo da linha de comboio do Tua, assim como o do troço do Pinhão ao Pocinho, na linha do Douro.

33	, assim como a	4	0.009

Resultados da Validação

Autor A - José Pacheco Pereira

É por causa de Manuela Ferreira Leite, do PSD, de Manuel Alegre, do BE, do PCP e acima de tudo por causa de José Sócrates, e do PS entre Alberto Martins e Vitalino Canas.

30	acima de tudo ,	5	0.007

É isso que significa a credibilidade, palavra com muito mais conteúdo do que parece e que muda muito mais coisas do que se imagina

192	o que significa que	3	0.004

Resultados da Validação
Autor B - António Barreto

Mas a verdade é que a alegada cornucópia é muito inferior ao necessário.

1	a verdade é que	14	0.033

Ao mesmo tempo que se ouvem declarações messiânicas sobre as novas fontes de energia e a poupança de combustíveis, anunciam-se mais auto-estradas, pontes e viadutos.

15	ao mesmo tempo ,	6	0.014

Prepara-se o fecho definitivo da linha de comboio do Tua, assim como o do troço do Pinhão ao Pocinho, na linha do Douro.

33	, assim como a	4	0.009

Conclusão

- Existem diferenças semânticas significativas?
- Como poderemos interpretar os dados?
- Os dados obtidos representam marcadores de autoria?
- A frequência de *hapax legomena* e *hapax dislegomena* será significativa?

Rui Silva
rmsilva@me.com

Faculdade de Letras
Universidade do Porto