Riqueza lexical como critério de detecção de autoria

Rui Sousa Silva
Centro de Linguística da Universidade do Porto
Centre for Forensic Linguistics, Aston University

Abstract

Recent developments in forensic discourse analysis (i.e., forensic linguistics) enabled authorship studies and authorship recognition to be more reliable in determining the real author in cases of plagiarism or even criminal offense where linguistic proof is involved. In this paper, we discuss the usefulness of discourse markers such as word length, sentence length and lexical density in authorship recognition in Portuguese. We show that these markers remain valid in Portuguese, as in English, and that texts written by different authors use distinct linguistic patterns, and show different authorship markers that differentiate them from all other texts.

Keywords: Authorship, Plagiarism, Lexical Richness, Forensic Linguistics, Discourse Analysis
Palavras-chave: Autoria, Plágio, Riqueza Lexical, Linguística Forense, Análise de Discurso

1. Da análise de discurso à linguística forense

Nos últimos anos, diversos autores têm proposto diferentes teorias e abordagens ao estudo e análise do discurso (Coulthard, 1977; Dijk, 1997; Fairclough & Wodak, 1997; Sinclair, 1991). Embora algumas destas teorias analisem a interacção entre o discurso e a sociedade (Dijk, 1997; Fairclough & Wodak, 1997) como forma de revelar, por exemplo, questões de ideologia dominante ou desvendar relações de poder expressas através da linguagem, outras procuram analisar sobretudo o discurso enquanto realização linguística (Coulthard, 1977; Sinclair, 1991) ou, inclusivamente, enquanto relação entre a linguística e a lei como forma de linguística forense (isto é, análise forense do discurso) (Coulthard & Johnson, 2007). É neste último, o contexto da linguística forense, que se insere a presente abordagem aos estudos de autoria.

Num sentido restrito, conforme refere Gibbons (2003: 12), a linguística forense constitui-se enquanto elemento de prova linguística utilizada na e pela lei, por exemplo enquanto prova pericial. Num sentido mais lato, como explica aquele autor, a linguística forense aplica conhecimentos especializados em linguística a casos legais. Porém, em virtude da diversificação das aplicações da linguística forense, o termo utiliza-se, actualmente, num sentido genérico como sinónimo de lei e linguagem (cf., por exemplo,

***Textos Seleccionados. XXIV Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2009, pp. 575-587**

Gibbons (2003)), incluindo: (a) a identificação de autoria; (b) a identificação modal; (c) a tradução e interpretação jurídica; (d) a transcrição de declarações e depoimentos; (e) a linguagem e o discurso dos tribunais; (f) os direitos linguísticos; (g) a análise de depoimentos e interrogatórios policiais; (h) a fonética forense e (i) o estatuto textual.

No presente artigo centramo-nos na primeira destas aplicações (a identificação de autoria), uma vez que a investigação nesta área permite determinar o verdadeiro autor em casos de disputa de autoria e, consequentemente, utilizar elementos linguísticos como prova na investigação de ocorrências de plágio ou ajudar a descobrir o verdadeiro autor de mensagens de resgate ou de ameaças num contexto de investigação policial. O enfoque nesta vertente da linguística forense justifica, por conseguinte, a adopção de uma metodologia orientada para as questões de identificação de autoria assente na aplicação de alguns marcadores de autoria idiossincráticos, como sejam a dimensão média dos textos ou das frases, questões de estilística forense (McMenamin, 2002), a frequência de determinados padrões linguísticos, o estudo das ocorrências de *hapax legomena* e *hapax dislegomena*, entre outros. Iremos, assim, analisar a validade de alguns destes marcadores no reconhecimento de autoria de textos de dois autores distintos.

2. Autoria e linguística forense

2.1. Autoria

O estudo das questões de autoria, ainda que actual, é um tema debatido há vários séculos, nomeadamente na literatura (Hänlein, 1998). Na Grécia Antiga, a autoria e a autoridade das fontes constituía o garante da credibilidade de um texto. Porém, os conceitos modernos de indivíduo, de autoria e de autor individual, com origem na filosofia humanista do Renascimento, vieram alterar esta perspectiva. Com a invenção da imprensa, atribuída a Gutenberg, no século XV, alteraram-se os cânones textuais tradicionais e os textos passaram a ser impressos em série, adquirindo um cariz económico para os seus autores. No século XVI, em virtude do desenvolvimento da escrita enquanto profissão e do declínio do patronato na literatura, as questões económicas assumiram um papel mais preponderante na escrita e aumentaram as preocupações com o usufruto, a cópia e, inclusivamente, o roubo da "palavra escrita", de tal modo que, em 1624, o Bispo Richard Montagu terá utilizado metaforicamente a palavra "plágio" (do Latim *plagiarius* ("raptor") e do Grego *plagion* ("rapto")) num contexto literário, permitindo que as palavras começassem a ser vistas como propriedade. No início do século XVIII foi criada no Reino Unido a primeira legislação destinada a gerir casos de plágio e de violação dos direitos de autor, passando os escritores e os editores a usufruir de direitos legais sobre os seus textos. Com o Neoclassicismo, reconsiderou-se o conceito de imitação, passando a dar-se um novo destaque à originalidade na escrita. Em meados do século XVIII, escritores como Alexander Pope e Samuel Johnson alertavam para casos de plágio e surgia publicamente o conceito de propriedade de bens considerados morais (isto é, não materiais). A nível mundial, este conceito e as políticas legais de autoria com ele relacionadas adquiriram maiores proporções com a criação e regulamentação de normas e acordos internacionais de direitos de autor (como, por exemplo, a Convenção de Berna),

através dos quais os autores passaram a ver os seus direitos protegidos, não só no seu próprio país, mas também em todos os países signatários da Convenção. Desde então, o drástico aumento da produção literária levou a questionar a propriedade de um bem comum, socialmente partilhável e partilhado, como é o caso das palavras (Angèlil-Carter, 2000; Howard, 1995; Scollon, 1994; 1995), bem como o conceito de originalidade[1], que passou a confundir-se frequentemente com estilo, na literatura.

Um dos problemas com que se depara a análise de autoria consiste na utilização de um bem comum como a língua e os mecanismos linguísticos com ela relacionados (e, portanto, utilizado de forma idêntica pelos falantes dessa língua) como factor de identificação de marcas pessoais. O peso deste argumento aumenta consideravelmente se considerarmos que, em interacção social, a língua tende a moldar-se a um registo e a acomodar-se às circunstâncias de utilização (Trudgill, 1974). No entanto, é comum os leitores assíduos de determinados autores aperceberem-se de algo familiar quando lêem textos de estilo idêntico, suscitando aquilo que Hänlein (1998) designa sensação de "déjà--vu". Esta sensação será mais acentuada quanto mais marcado (isto é, original) for o estilo do autor. Esta definição assenta no princípio de que, ao lermos um texto ou um excerto de determinado autor, se estivermos familiarizados com o estilo desse autor conseguiremos identificá-lo, ainda que de forma meramente intuitiva. Este processo, que permite, não raras vezes, detectar tentativas de imitação, designa-se reconhecimento de autoria. Porém, o princípio de estilo como elemento idiossincrático de um autor não é novo. No século XVIII, Buffon (19--) justificava-o através da premissa de que "Le style est l'homme même"[2], ou seja, cada falante de uma língua possui um idiolecto próprio, uma forma única de escrever ou de falar, que o distingue dos demais falantes da mesma língua. Assentando neste conceito de estilo, é frequente estabelecer-se um paralelo entre a estilística literária e a estilística forense, uma vez que (a) ambas procuram determinar aspectos de estilo como metodologia de reconhecimento do estilo do autor e (b) os conceitos de "estilo individual" (utilizado pela primeira) e de "idiolecto" (utilizado pela segunda) são parcialmente coincidentes (Hänlein, 1998). Porém, as duas vertentes da estilística possuem âmbitos e objectivos diferentes, uma vez que o idiolecto assenta em escolhas determinadas pela interacção social. Na linguística contemporânea, a noção de estilo individual é constituída, não só pelo "comportamento linguístico" de autores conhecidos, mas também pelas escolhas estilísticas do falante de uma língua.

Porém, com o aumento drástico do volume de produção literária, torna-se cada vez mais difícil garantir a consagração do direito de autoria individual, não sendo possível determinar a propriedade da multiplicidade de conjugações possíveis de diferentes palavras de outra forma que não recorrendo à análise do estilo do autor.

---

[1] A legislação portuguesa, nomeadamente o Código do Direito de Autor e dos Direitos Conexos, considera como obras originais as "criações intelectuais do domínio literário, científico e artístico, quaisquer que sejam o género, a forma de expressão, o mérito, o modo de comunicação e o objectivo", que não sejam confundíveis com qualquer outra obra do mesmo género de outro autor anteriormente divulgada ou publicada.

[2] "O estilo é o próprio Homem".

2.2. Linguística forense

Pelas suas características intrínsecas e pela natureza que assume em casos legais, o reconhecimento de autoria enquanto ramo especializado da linguística aplicada necessitou de se desenvolver num sentido diametralmente oposto ao da estilística literária para poder constituir-se como elemento inerentemente científico e, consequentemente, válido em contextos jurídicos. Revelando-se útil em casos como a disputa de autoria, o reconhecimento de autoria inscreve-se, na linguística, no domínio da linguística forense, isto é, no ramo da linguística aplicada que consiste em utilizar os resultados da sua investigação no julgamento de actividades legais e jurídicas de recolha e interpretação de provas linguísticas. Com uma natureza sociolinguística, o reconhecimento de autoria em linguística forense integra-se, por conseguinte, no ramo mais lato da análise de discurso.

A análise de discurso, que teve origem na linguística descritiva enquanto forma de análise (sociolinguística) de realização linguística, com a publicação de "An Introduction to Discourse Analysis" (Coulthard, 1977), foi objecto de diversas abordagens desde então, assumindo uma dimensão cada vez mais sociológica como análise da interacção entre o discurso e a sociedade (Dijk, 1997; Fairclough & Wodak, 1997). Entretanto, desdobrou-se em novos ramos, através da junção de adjectivos ao termo inicial "análise de discurso" como, por exemplo, "análise crítica do discurso" e "análise crítica feminista do discurso" e "análise forense do discurso" (Coulthard & Johnson, 2007) – como se designa frequentemente a linguística forense. Este é o ramo da análise de discurso que procura aplicar a investigação em áreas da sociolinguística como a dialectologia, a variação linguística e a estilística, entre outras, e na qual se inscreve o reconhecimento de autoria.

Os desenvolvimentos tecnológicos dos últimos anos, nomeadamente o desenvolvimento de ferramentas e métodos de análise de *corpora*, permitiram à linguística aplicada desenvolver-se em novos sentidos, possibilitando a gestão e processamento de grandes quantidades de texto, de outro modo impensável, como forma de resolução de problemas (socio)linguísticos do "mundo real". Paralelamente, a linguística aplicada viu-se obrigada a desenvolver-se em novos sentidos para poder aplicar os resultados da investigação em linguística à resolução de problemas concretos. Um destes exemplos é a abordagem ao plágio: se, por um lado, o desenvolvimento dos motores de busca e da Internet em geral permitiram recolher e copiar informação mais fácil, rápida e simplesmente, por outro lado também permitiu que esses casos passassem a ser detectados com maior facilidade.

3. A linguística forense e a demonstração de resultados

Em linguística forense, aquilo que conduz, muitas vezes, à análise e investigação de um texto suspeito para determinar o seu verdadeiro autor é a intuição, a sensação de "déjà-vu", que se baseia em opiniões e impressões pessoais, sendo, por isso, um método extraordinariamente subjectivo. Em casos de plágio académico, por exemplo, a sensação de que um texto se encontra "demasiadamente bem escrito" levanta suspeitas relativamente à sua autoria. Porém, para que as suspeitas possam ser confirmadas, é necessário proceder a uma análise baseada em dados científicos concretos, válidos e

fiáveis que permitam justificar objectivamente essa suspeita. Só então será possível confrontar a pessoa responsável pelo crime de usurpação de propriedade intelectual e proceder, se for caso disso, à aplicação das correspondentes medidas de acção disciplinar. No caso de situações de violação dos direitos de autor punível por lei, ao abrigo do Código do Direito de Autor e dos Direitos Conexos, a situação é ainda mais complexa: as provas utilizadas para lançar a acusação têm que ser sólidas e bem fundamentadas para que, por um lado, sejam aceites em tribunal e, por outro lado, não originem contra--acusações e processos legais por difamação.

Neste contexto, o linguista forense desempenha um papel fundamental como perito em tribunal para demonstrar, através de métodos científicos válidos, quem é o autor de determinado texto. Deve, por isso, socorrer-se da análise estatística de aspectos linguísticos do texto, identificando as constantes e as variantes de estilo (ou seja, as regularidades e as irregularidades da escrita do autor), com base em marcadores de discurso, entre os quais se incluem a dimensão média dos textos, o comprimento médio das frases e das palavras, a frequência de determinados padrões linguísticos, as ocorrências de *hapax legomena* e *hapax dislegomena* e a riqueza lexical.

4. Objectivos e metodologia

Neste estudo, analisamos as ocorrências de unidades lexicais de dois autores com o objectivo de averiguar, utilizando padrões de diversidade e abrangência, se a riqueza lexical, um marcador de discurso utilizado em análise de autoria pela linguística forense inglesa (Coulthard & Johnson, 2007; Woolls & Coulthard, 1998), também é válida para determinar a autoria em português. Assim, recorremos aos estudos em linguística de *corpus* (McEnery & Wilson, 1996) para criar um *corpus* com cerca de 100.000 palavras dos textos escritos pelos cronistas António Barreto e José Pacheco Pereira, publicados no jornal *Público* entre Janeiro e Dezembro de 2007. Com o objectivo de proceder a uma análise estatística dos textos, nomeadamente em termos de rácio do tipo de palavras relativamente ao total de átomos do texto (TTR), comprimento médio das palavras, comprimento médio das frases utilizadas e riqueza lexical, utilizamos duas ferramentas de gestão de *corpora*: o *Corpógrafo* (Sarmento, Maia & Santos, 2004) e o *Wordsmith Tools* (Scott, 2008). Para avaliar a validade e a fiabilidade das medidas de reconhecimento de autoria discutidas no presente artigo, procedemos, então, à "análise cega" de dois textos da autoria destes dois autores, publicados no mesmo jornal (*Público*) em Janeiro de 2008. A estes dois textos atribuímos aleatoriamente as designações de "Autor A" e "Autor B".

4.1. Rácio tipo de palavras/total de átomos (TTR)

O rácio do tipo de palavras relativamente ao total de átomos do texto (TTR, ou *type/token ratio*) constitui um método tradicional utilizado em estudos de autoria para determinar a diversidade vocabular de um autor. Partindo do princípio de que um texto produzido naturalmente é constituído por palavras cuja maioria é utilizada várias vezes no texto e apenas uma parte é diferente, este método consiste em calcular a percentagem de tipos de palavras face ao total de átomos do texto. Considerando, por exemplo, que um

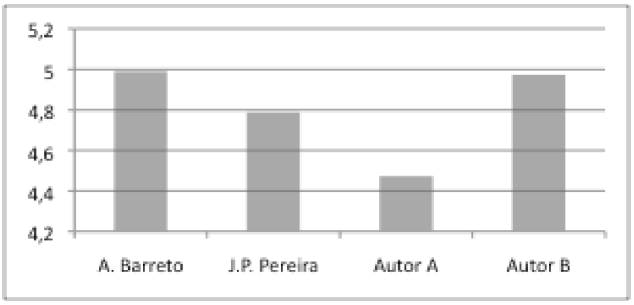
texto apresenta dez ocorrências da preposição "de", teremos dez átomos correspondentes à preposição "de", mas apenas um tipo – o que significaria um rácio de 10%. De acordo com este método, a um TTR mais elevado corresponde uma utilização mais diversificada de vocabulário.

Este modelo tem, no entanto, sido questionado. Por um lado, ferramentas como o *Wordsmith Tools* consideram átomos apenas palavras e não todos os elementos do texto, como a pontuação; o *Corpógrafo*, pelo contrário, considera a pontuação como átomos. Esta distinção é fundamental, uma vez que a pontuação constitui um marcador de estilo essencial. Por outro lado, os testes realizados mostram que este método não é fiável, uma vez que o valor de autoria obtido é relativo e varia consoante o tamanho dos textos. A solução proposta pelo *Wordsmith Tools* para ultrapassar este problema consistiu em basear a análise no rácio do número de tipo de palavras relativamente ao total de átomos padrão (STTR, "standardised type/token ratio"), ou seja, aos valores médios calculados por cada conjunto de mil palavras. Mais uma vez, esta medida não pode ser considerada, uma vez que os textos apresentam dimensões diferentes, muitos deles com menos de mil palavras. Consequentemente, neste estudo não incluímos a análise deste marcador como marcador válido de reconhecimento de autoria.

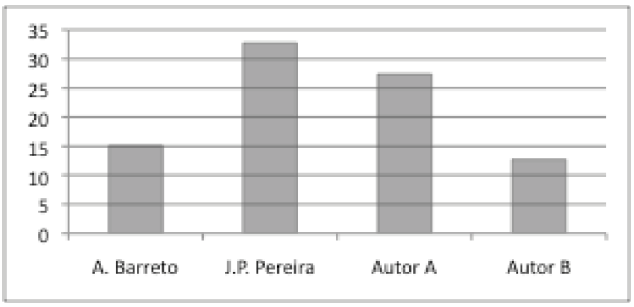
4.2. Comprimento médio das palavras e das frases utilizadas

Estudos de alguns autores (Coulthard & Johnson, 2007; Woolls & Coulthard, 1998) mostram que o comprimento médio das palavras utilizadas (isto é, o número médio de caracteres por palavra) e o comprimento médio das frases utilizadas (que consiste no número de palavras por frase) são elementos linguísticos muito variáveis de autor para autor, ou de acordo com o tipo de escrita. Tomando como exemplo o comprimento médio das palavras, constatamos, com Hänlein (1998), que palavras mais longas estão, normalmente, associadas a uma escrita mais formal e/ou a textos escritos, enquanto a utilização de palavras mais curtas se encontra relacionada com uma escrita mais informal e/ou textos orais. No entanto, uma vez que diferentes autores apresentam preferências diferentes, sobretudo em termos de comprimento das frases, estes marcadores linguísticos possuem capacidade discriminatória, constituindo-se, por isso, como marcadores fiáveis e válidos na análise de autoria. Coulthard & Johnson (2007) defendem, ainda, que o comprimento médio das frases é um marcador de discurso particularmente significativo, uma vez que se encontra, normalmente, sob o controlo (sub)consciente do autor; efectivamente, é ao autor que compete decidir onde colocar a pontuação e, assim, continuar ou terminar a frase.

Nos quadros 1 e 2 apresentamos uma comparação do comprimento médio das palavras e das frases dos textos dos dois autores (quer nos textos cuja autoria conhecemos, quer nos textos de "teste cego"). Em ambos os casos, os resultados apontam para uma primeira hipótese de que o Autor B se encontra mais próximo de António Barreto, enquanto o Autor A se aproxima de José Pacheco Pereira.



Quadro 1: Comprimento médio das palavras



Quadro 2: Comprimento médio das frases

4.3. Riqueza lexical

A riqueza lexical (ou densidade lexical) é outro marcador de discurso que, segundo Coulthard e Johnson (2007), se encontra sob o controlo (sub)consciente do autor, a exemplo do que acontece com o comprimento médio das frases, assumindo, assim, particular relevância. A riqueza lexical baseia-se no conceito de *Riqueza* (vocabular) de Honoré (1979), que propõe uma unidade de medida da utilização de *hapax legomena* (palavras utilizadas uma única vez em determinado texto), com base no princípio de que, quanto maior for a proporção de palavras utilizadas uma única vez, maior será a riqueza vocabular.

A fórmula proposta por Honoré é:

$$100*logN/(1—V_1/V)$$

em que N corresponde à dimensão total do texto em palavras, V_1 corresponde ao vocabulário utilizado apenas uma única vez (*hapax legomena*) e V corresponde ao total de vocabulário no texto (tipos). Nesta fórmula, considera-se o vocabulário de forma genérica, incluindo todas as palavras do texto, sem fazer uma distinção entre elementos lexicais e elementos gramaticais. Este método constitui, por isso, uma medida de riqueza vocabular e não uma medida de riqueza lexical.

A *riqueza lexical* constitui uma unidade de medida proposta por Woolls (Coulthard & Johnson, 2007; Woolls & Coulthard, 1998) e assenta no princípio de que, enquanto o léxico transporta consigo o sentido, os elementos gramaticais servem de elo de ligação entre os elementos lexicais. A riqueza lexical corresponde, assim, aos itens lexicais que ocorrem uma única vez no texto, isto é, à diversidade de elementos lexicais utilizados, e demonstra-se estatisticamente em relação ao tamanho do texto. A um aumento do número de elementos lexicais corresponderá, assim, um aumento da riqueza lexical do texto. Esta proposta consiste em substituir simplesmente V_1 por LV_1, passando a considerar apenas para efeitos de riqueza linguística os *hapaxes* de natureza lexical e não todo o vocabulário. A fórmula resultante desta proposta é:

$$100*logN/(1—LV_1/V)$$

Esta alteração é justificada pelos autores (Woolls & Coulthard, 1998) como uma tentativa de contrabalançar o efeito de diferentes taxas de crescimento em textos de dimensões diferentes. Uma vez que a língua possui um conjunto relativamente pequeno de elementos gramaticais utilizados, consequentemente, com maior frequência e uma maior tendência de repetição, quanto mais longo for o texto, menor será a proporção de itens gramaticais utilizados uma única vez – e, logo, menor será a incidência de *hapaxes* sobre este conjunto de palavras. No caso da fórmula proposta por Honoré, os conjuntos "abertos" e os conjuntos "fechados" de palavras (i.e., as unidades lexicais e as unidades gramaticais) são misturados, originando um efeito de desproporcionalidade. Baseando-se em estudos anteriores que mostravam que, em textos de dimensões superiores a 500 palavras, o número de elementos gramaticais utilizados uma única vez no texto tendia a manter-se dentro de um intervalo limitado de 35 a 50 elementos, independentemente das dimensões do texto, enquanto o número de elementos lexicais continuava a aumentar proporcionalmente às dimensões do texto, aqueles autores propuseram a substituição da fórmula de riqueza vocabular pela fórmula de *riqueza lexical*, com o objectivo de permitir efectuar uma comparação fiável de textos de diferentes tamanhos escritos por diferentes autores.

Neste estudo, para determinar a riqueza das palavras utilizadas pelos dois autores, adoptámos o método de riqueza lexical proposto por Winter e Woolls (citado em Coulthard & Johnson, 2007), introduzindo, no entanto, uma ligeira alteração: a análise dos itens lexicais é efectuada recorrendo, em primeiro lugar, a um processo de lematização, de modo a considerar as formas lexicais não flexionadas como sendo uma mesma palavra e não palavras independentes. Embora Woolls e Coulthard (1998: 51) defendam que não deve aplicar-se a lematização em textos de pequenas dimensões e que as diferentes formas das palavras devem ser separadas e incluídas na lista de *hapax legomena*, julgamos, por

um lado, que a gramática da língua portuguesa, na medida em que permite uma flexão alargada, iria aumentar exponencialmente o número de *hapaxes* e, por outro lado, que os textos em análise são suficientemente longos para permitir uma análise estatística deste tipo com resultados significativos.

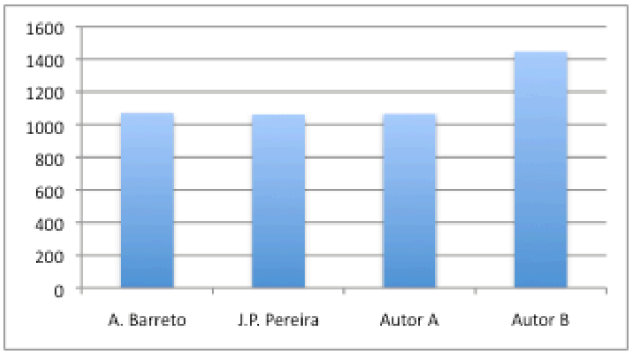
Em segundo lugar, consideramos que a pontuação constitui um factor muito importante no reconhecimento de autoria e não pode, por conseguinte, ser ignorada. No quadro 3 apresentamos os valores comparativos entre o número total de palavras (N) e o número total de átomos incluindo a pontuação (N1) dos textos, o que nos permite constatar que o total de átomos, incluindo a pontuação, não é um factor irrelevante:

	A. Barreto	J.P. Pereira	Autor A	Autor B
N (átomos)	34125	54792	1452	965
N1 (átomos: Corpógrafo)	41321	66032	1675	1172

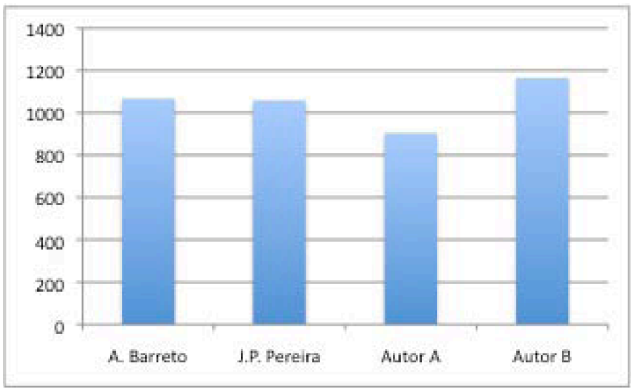
Quadro 3: Comparação de N (total de palavras) e N1 (total de átomos, incluindo pontuação)

No entanto, quer a fórmula proposta por Honoré, quer a fórmula proposta por Winter e Woolls, consideram como total de átomos o total de vocábulos do texto, não incluindo pontuação. Optámos, por isso, por utilizar a fórmula proposta por Winter e Woolls, substituindo o valor de N (número total de vocábulos) por N1 (número total de átomos no texto). A ferramenta que utilizamos para fazer este cálculo foi o *Corpógrafo* por permitir, contrariamente ao *Wordsmith Tools*, contabilizar como átomos todos os elementos do texto e não apenas o vocabulário.

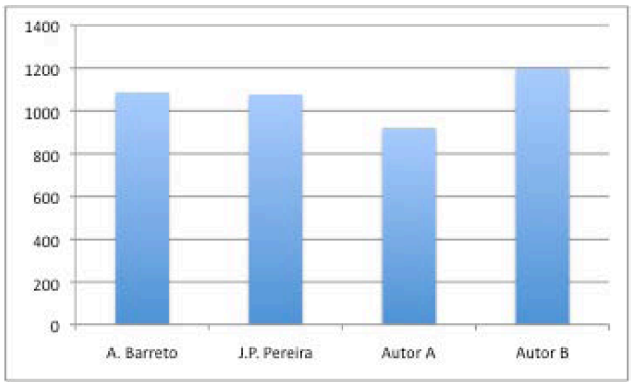
Os resultados da análise da riqueza vocabular, da riqueza lexical (considerando o vocabulário) e da riqueza lexical (considerando o número total de átomos dos textos) correspondentes aos dois autores são os apresentados nos quadros 4, 5 e 6:



Quadro 4: Riqueza vocabular (Honoré): *100 x log N / (1-(V1/V))*



Quadro 5: Riqueza lexical: Winter & Woolls: *100 x log N / (1-(LV1/V))*



Quadro 6: Riqueza lexical: *Corpógrafo*: *100 x log N1 / (1-(LV1/V))*

5. Discussão de resultados

A noção de idiolecto assenta no princípio de que os falantes de uma língua vão construindo, ao longo da sua vida, um vocabulário mais ou menos diversificado, conforme o falante, distinto do vocabulário dos demais falantes da mesma língua, apesar de construído sob circunstâncias idênticas (Bloch, 1948; Coulthard & Johnson, 2007). Estas diferenças revelam-se na realização linguística do quotidiano, uma vez que, por um lado, a gama de vocabulário que cada falante tem à sua disposição limita as opções linguísticas e, por outro lado, revela as preferências na selecção de determinados elementos em detrimento de outros. Esta co-selecção de elementos linguísticos constitui, por isso, uma co-selecção individualizada e, por conseguinte, diferente das demais. Este processo,

apesar de muitas vezes referido metaforicamente como "impressão digital" linguística (Coulthard & Johnson, 2007; Hänlein, 1998), não proporciona, no entanto, a mesma precisão que a lofoscopia em identificação de impressões digitais, uma vez que, por um lado, a possibilidade de criar uma imensa base de dados com informações de todos os falantes de uma língua é ínfima, se não impossível, e, por outro lado, qualquer amostra linguística de um falante poderá ser apenas parcial, não contendo todas as informações necessárias com vista à identificação de uma pessoa. O reconhecimento da autoria baseia-se, sobretudo, na comparação de amostras linguísticas com outras amostras disponíveis, normalmente em número limitado. Daí que seja necessário determinar alguns marcadores válidos e fiáveis para realizar com êxito esse reconhecimento de autoria.

Neste estudo, o primeiro marcador de discurso que analisámos foi o comprimento médio das palavras. Os dados apresentados (Quadro 1: Comprimento médio das palavras) mostraram que, à luz do *corpus* utilizado, o autor António Barreto utiliza palavras com uma média de 4,99 caracteres por palavra, comparativamente a 4,78 palavras nos textos de José Pacheco Pereira. No caso dos textos "cegos", o Autor A apresenta uma média de 4,47 caracteres por palavra, comparativamente a 4,97 caracteres por palavra do Autor B. Apesar de a média do Autor B se encontrar entre a média dos autores José Pacheco Pereira e António Barreto, a aproximação à média do autor António Barreto é particularmente marcada (quase coincidente, com uma diferença de apenas 0,02 pontos).

Em relação ao comprimento médio das frases, as diferenças entre os textos de António Barreto e José Pacheco Pereira e entre o Autor A e o Autor B são significativas: a média de José Pacheco Pereira (32,75 palavras por frase) encontra-se muito próxima da média do Autor A (27,34 palavras por frase), enquanto a média de António Barreto (15,13 palavras por frase) se aproxima mais da média do Autor B (12,69 palavras por frase). Socorremo-nos de Coulthard e Johnson (2007: 165) para realçar que o comprimento médio das frases, encontrando-se, normalmente, sob o domínio (sub)consciente do autor, constitui um marcador de estilo relevante. Estes dados estatísticos são, ainda, mais relevantes se considerarmos que a diferença apurada entre António Barreto e José Pacheco Pereira (17,62) se encontra praticamente em paralelo com a diferença apurada entre o Autor A e o Autor B, respectivamente (14,65). Por outro lado, a diferença entre os textos de José Pacheco Pereira e o Autor A (5,41) e António Barreto e o Autor B (2,44) parece confirmar, clara e inequivocamente, os dados apurados com o estudo do comprimento médio das palavras.

Na análise da riqueza lexical, os resultados indicam, igualmente, uma identificação de António Barreto com o Autor B e de José Pacheco Pereira com o Autor A, conforme consta dos quadros apresentados acima (Quadro 4: Riqueza vocabular (Honoré): *100 x log N / (1-(V1/V))*, Quadro 5: Riqueza lexical: Winter & Woolls: *100 x log N / (1-(LV1/V))* e Quadro 6: Riqueza lexical: *Corpógrafo*: *100 x log N1 / (1-(LV1/V))*). Nos três casos analisados (riqueza vocabular segundo a fórmula proposta por Honoré, riqueza lexical baseada na proposta de Honoré, com as devidas adaptações de Winter e Woolls, e riqueza lexical baseada no *Corpógrafo*), o Autor B distancia-se sempre dos restantes autores em termos de riqueza lexical; porém, apesar deste relativo distanciamento, a aproximação é sempre maior ao autor António Barreto do que ao autor José Pacheco Pereira. O Autor A, por outro lado, encontra-se mais próximo do autor José Pacheco Pereira do que do autor António Barreto nas três análises de riqueza lexical realizadas.

Os testes realizados em termos de comprimento médio das palavras, comprimento médio das frases e riqueza lexical (baseada em três fórmulas de cálculo diferentes) apontam, neste caso de reconhecimento de autoria, para resultados estatísticos inequívocos, segundo os quais o Autor B é António Barreto e o Autor A é José Pacheco Pereira. Os dados do "teste cego" confirmam esta hipótese.

Relativamente à riqueza lexical, constatámos que a fórmula *100 x log N1 / (1- -(LV1/V))*, proposta em termos de análise utilizando o *Corpógrafo*, testa mais exaustivamente o reconhecimento de autoria do texto, uma vez que inclui a análise da pontuação. Os valores de riqueza lexical apresentam, assim, um peso relativo inferior, sendo necessária uma maior representatividade para serem significativos.

6. Conclusão

Os resultados das várias análises realizadas mostram que os três marcadores testados apresentaram um desempenho excelente e consistente em todos os testes, o que nos permite reconhecê-los como factores válidos e fiáveis de reconhecimento de autoria em português, a exemplo do que acontece em inglês. Os resultados dos "testes cegos" realizados correspondem à tendência que se verifica nos valores obtidos através do estudo do *corpus* de referência (isto é, os textos publicados pelos dois autores durante o ano de 2007 no jornal *Público*).

Não poderemos, também, deixar de considerar, como defende Hänlein (1998: 57), que a política editorial do órgão de comunicação social onde são publicados os textos pode exercer um elevado grau de influência no sentido da harmonização dos marcadores estilísticos de autoria, de acordo com o respectivo livro de estilo. No caso em apreço, os textos dos dois autores foram publicados no mesmo jornal e encontram-se, por conseguinte, sujeitos às mesmas normas de revisão e edição, constituindo, assim, um factor potencialmente influenciador do estilo de autoria. Porém, a confirmar-se, esta influência exerceria um efeito inverso, ou seja, contribuiria para uma aproximação e não para um afastamento estilístico. Este factor confere aos testes realizados uma importância ainda maior, pois, conforme esta análise permite constatar, os textos dos dois autores diferentes utilizam, efectivamente, padrões linguísticos distintos e idiossincráticos; para além de diferentes estruturas e dimensões das frases e dos textos (factores que, como leitores comuns, poderemos apreender intuitivamente), cada um dos autores analisados possui um idiolecto próprio, com marcas de autoria distintas, que permitem diferenciar essa produção linguística das demais.

Apesar de apresentarem resultados inequívocos, os métodos de análise estudados possuem uma utilidade limitada pois, embora funcionem no reconhecimento de autoria de textos com *corpus* de referência (isto é, para comparar textos de autoria desconhecida com um conjunto de textos cuja autoria é conhecida), não são suficientes para determinar o autor de um texto entre um universo ilimitado de autores. Constatamos, portanto, com Coulthard e Johnson (2007), que não poderemos aplicar a "lofoscopia linguística" para determinar a autoria de um texto. Porém, a aplicação dos testes realizados no âmbito deste estudo utilizando um *corpus* de referência permite reconhecer a autoria de textos, ainda que relativamente pequenos (iguais ou inferiores a mil palavras).

Consideramos, finalmente, que esta metodologia de análise serve, ainda, de ponto de partida para o desenvolvimento de investigação futura no domínio da autoria, nomeadamente em identificação de situações de plágio.

Referências

Angèlil-Carter, S. (2000) *Stolen language?: plagiarism in writing*. Harlow: Longman.

Bloch, B. (1948) A set os postulates for phonemic analysis. *Language* 24, pp. 3-46.

Buffon, M. D. (19--) *Discours sur le style prononcé a l'académie française*. Paris: Librairie Classique Eugène Belin.

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

Coulthard, M., & A. l. Johnson (2007) *An Introduction to Forensic Linguistics: Language in Evidence*. London and New York: Routledge.

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

Fairclough, N. & R. Wodak (1997) Critical Discourse Analysis. In T. A. v. Dijk (ed.) *Discourse Studies: A Multidisciplinary Introduction – Discourse as Social Interaction* (Vol. 2, pp. 258-284). Londres: SAGE Publications Ltd.

Gibbons, J. (ed.) (2003) *Language and the Law*. London: Longman.

Hänlein, H. (1998) *Studies in Authorship Recognition – A Corpus-based Approach* Frankfurt: Peter Lang.

Honoré, A. (1979) Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin* 7 (2), pp. 172-177.

Howard, R. (1995). Plagiarisms, Authorships, and the Academic Death Penalty. *College English* 57 (7), pp. 788-806.

McEnery, T. & A. Wilson (1996) *Corpus Linguistics: An Introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

McMenamin, G. R. (2002) *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton and New York: CRC Press.

Sarmento, L., B. Maia & D. Santos (2004) *The Corpógrafo – a Web-based environment for corpora research*.

Scollon, R. (1994) As a matter of fact: The changing ideology of authorship and responsibility in discourse. *World Englishes* 13 (1), pp. 33-46.

Scollon, R. (1995) Plagiarism and ideology: Identity in intercultural discourse. *Language in Society* 24, pp. 1-28.

Scott, M. (2008) Wordsmith Tools (Version 5). Oxford: Oxford University Press.

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

Trudgill, P. (1974) *The social differentiation of English in Norwich*. London: Cambridge University Press.

Woolls, D., & M. Coulthard (1998) Tools for the Trade. *International Journal of Speech, Language and the Law* 5 (1), pp. 33-57.