Universidade de Lisboa
Faculdade de Letras

ATRIBUIÇÃO DE AUTORIA EM LINGUÍSTICA FORENSE:
UMA ANÁLISE COMBINADA PARA IDENTIFICAÇÃO DE AUTOR ATRAVÉS DO TEXTO

Liliana Rita de Amorim Romão Teles

Trabalho final orientado pela Prof.ª Doutora Rita Marquilhas
especialmente elaborado para a obtenção do grau de mestre em Linguística

Dissertação
Mestrado em Linguística
2015

Universidade de Lisboa
Faculdade de Letras

ATRIBUIÇÃO DE AUTORIA EM LINGUÍSTICA FORENSE: UMA ANÁLISE COMBINADA PARA IDENTIFICAÇÃO DE AUTOR ATRAVÉS DO TEXTO

Liliana Rita de Amorim Romão Teles

Trabalho final orientado pela Prof.ª Doutora Rita Marquilhas
especialmente elaborado para a obtenção do grau de mestre em Linguística

Dissertação
Mestrado em Linguística
2015

Agradecimentos

Esta dissertação é o resultado de longos meses de trabalho que apenas foram possíveis devido ao apoio de algumas pessoas.

Em primeiro lugar, agradeço à Professora Doutora Rita Marquilhas pela orientação, prontidão e pelas observações práticas e construtivas durante todo este processo. Agradeço também o apoio do Doutor João Silva, sem o qual não teria sido possível executar este trabalho nos moldes em que o imaginei de início, bem como a sua análise crítica e colaboração científica, sempre dispostas e esclarecedoras.

Agradeço também aos meus colegas do Centro de Linguística da Universidade de Lisboa pelo apoio, pelos contributos, revisões e discussão animada de ideias, bem como aos meus amigos que se mantiveram por perto, constantes, nos momentos mais complexos dos últimos meses.

Sou também muito grata pelo apoio da minha família, por contribuírem de forma ativa no reunir das condições práticas para a concretização deste trabalho, quando foi particularmente necessário.

O meu especial e sentido agradecimento ao Ruben, por tudo.

Para o Ruben.

"(...) Analisando-me à tarde, descubro que o meu sistema de estilo assenta em dois princípios, e imediatamente, e à boa maneira dos bons clássicos, erijo esses dois princípios em fundamentos gerais de todo estilo: dizer o que se sente exatamente como se sente – claramente, se é claro; obscuramente, se é obscuro; confusamente, se é confuso; compreender que a gramática é um instrumento, e não uma lei."

Bernardo Soares

Resumo

Com esta dissertação pretendemos verificar em que medida uma análise combinada, quantitativa e qualitativa, pode ser a abordagem adequada para casos forenses de atribuição de autoria a textos de valor probatório.

Assumindo que não é possível compreender a variedade linguística de um indivíduo sem ter previamente um conhecimento da variedade própria da comunidade em que este está inserido, partimos de um conceito de variação da língua a nível individual que não é propriamente o de idioleto, mas sim o de estilo idioletal, o conjunto das escolhas do falante individual a partir do sistema linguístico da sua própria comunidade (Labov (2006/1966):5, Turell (2010)).

Reunimos um *corpus* de cartas variado para verificar qual a possibilidade de atribuir o autor certo a um texto questionado. O *corpus* incluía um conjunto de 48 cartas redigidas anonimamente por doze informantes da mesma faixa etária e do mesmo dialeto, com controlo das variáveis "formação curricular" e "género". Para a análise quantitativa, usámos o classificador de uma Máquina de Vetores de Suporte, método usado frequentemente em estudos de atribuição de autoria, e verificámos a sua taxa de sucesso na atribuição de género, formação curricular e autoria a cada carta de ameaça, usando, como *corpus* textual de treino, as restantes cartas de cada autor. Numa segunda fase, repetimos o teste de classificação, mas apenas considerando uma carta de ameaça adicional, tomada como "TextoQ". Para a análise qualitativa, recolhemos os elementos linguísticos que se afiguravam reveladores do estilo do autor no texto questionado e procurámos identificar traços coincidentes num conjunto de cartas selecionadas do *corpus*.

Com esta experiência em ambiente controlado, conseguimos, recorrendo a uma análise combinada, atribuir a autoria de um texto hipoteticamente anónimo. Julgamos ter assim contribuído para a análise das diferenças nos enunciados escritos individuais, para a interpretação dos resultados do seu processamento computacional, e, consequentemente, para o avanço da linguística forense no contexto do estudo do português europeu.

Palavras-chave: variação, autoria**,** estilo idioletal, linguística forense, máquinas de vetores de suporte.

Abstract

With this dissertation we intend to verify in what way the combined analysis, both qualitative and quantitative, may be the suitable approach to forensic cases of authorship attribution to written texts to be used as instrumental proof.

Considering that is not possible to understand the linguistic variety of an individual without previously having the knowledge of his community's variety, we assume that the most adequate concept is not the one of an idiolect but rather the concept of idiolectal style, in the sense of the set of the speaker's choices in the linguistic system of his own community. (Labov (2006/1966), Turell (2010))

In order to check if it is possible to assign the right authorship to a given text, we collected a *corpus* with 48 letters written anonymously by 12 informants of the same age group and sharing the same dialect. We controlled the variables "educational curriculum" and "gender". For the quantitative analysis, we used a Support Vector Machine (SVM), as it is frequently used in the authorship attribution studies. Afterwards, we checked the success rate of the SVM classifier on the following tasks: authorship, educational curriculum and gender attribution for each of the threat letters, using the other letters from each author as a training *corpus*. In a second stage, we repeated the classification test, considering only an additional threat letter as a disputed text. In order to make the qualitative analysis, we gathered the features from the disputed text that could reveal the linguistic style of an unknown author. Finally, we matched those features with the selected letters from the sample *corpus.*

By running these tests on a controlled environment it was possible to make authorship attribution to a disputed text, using a combined analysis. Thus, we consider this dissertation as a contribution not only to the analysis of individual written discourse, but also to the interpretation of the results of its computational processing, and, finally, to the progress of forensic linguistics in European Portuguese.

Keywords: variation, authorship attribution, idiolectal style, forensic linguistics, support vector machines.

1 – Introdução

Atribuir autoria a um texto pode ser útil para diversas áreas de estudo, áreas como a Linguística, o Direito e a Crítica Textual, já que os textos são ao mesmo tempo instâncias de comportamento individual, gramatical, social e cultural. Daí que venha sendo pertinente, desde há séculos, responder à repetida pergunta "Quem escreveu este texto?". No entanto, as metodologias têm compreensivelmente evoluído, envolvendo na contemporaneidade contribuições de áreas tão distintas como a estatística e a aprendizagem automática.

A atribuição de autoria textual é, em sentido lato, a habilidade de inferir as caraterísticas de um autor a partir das caraterísticas dos documentos escritos por esse mesmo autor (Juola (2006: 233)). Mais concretamente, o típico problema de atribuição de autoria, e também o mais estudado, envolve atribuir a um dado texto questionado o seu respetivo autor, de entre um conjunto limitado de personagens possíveis. No entanto, também se considera abrangida pelos problemas de atribuição de autoria a tarefa de descobrir não apenas identidade individual, mas também identidade de grupo, advinda de factores de identificação tais como o género, o grau de formação curricular ou o dialeto (Juola 2006:299)).

A aplicação de métodos analíticos de base computacional ou estatística foi uma evolução natural do processo de atribuição de autoria, que assim foi incorporando mecanismos automáticos em busca da redução do impacto de eventuais erros humanos. Continuando com Juola (2006:272), encontramos aí a exposição de uma grande variedade de métodos disponíveis, dos analíticos não supervisionados aos analíticos supervisionados. Dentro dos métodos analíticos não supervisionados, isto é, que não necessitam de uma delimitação de traços pré-definida, encontramos a Análise em Componentes Principais (PCA), os Espaços Vetoriais, o Escalonamento Multidimensional (MDS) e a Análise de Clusters. Entre os métodos analíticos supervisionados, o autor descreve os métodos de estatística pura (ANOVA, t-test, etc..), a Análise Linear Discriminante (LDA), os métodos baseados em distância, os métodos básicos de aprendizagem automática, e, finalmente, as Máquinas de Vetores de Suporte (Support Vector Machines, SVMs). Estas últimas têm-se consolidado como método preferencial nos estudos de atribuição de autoria de base computacional, conforme Juola (2006:286) afirma, "SVMs generally outperform other methods of

classification such as decision trees, neural networks, and LDA — which in turn has been shown to outperform simple unsupervised techniques such as PCA". Contudo, e como sublinha o autor, esta conclusão não justifica sozinha a decisão de selecionar as SVMs como o método mais indicado. Importa enriquecer o elenco de experiências desenvolvidas com este e outros métodos antes de confirmar a vantagem do recurso às Máquinas de Vetores de Suporte.

Além do apuramento de um método de classificação automática textual, para assegurar as melhores práticas em atribuição de autoria em linguística forense, é também necessário reunir condições estruturais, como lembra Chaski (2013). Quer isto dizer que as metodologias apropriadas a este tipo de investigação envolvem um contexto de experiência independente de qualquer disputa legal, o uso de dados com variáveis controladas, o emprego de textos comparáveis aos dos casos judiciais reais, a inclusão de um protocolo experimental estabelecido empiricamente, o controlo de erros cumulativos, a possibilidade de replicação da experiência, e, finalmente, a adequada fundamentação em outras investigações e na teoria científica (Chaski (2013:336-344)).

Nesta dissertação, ao testarmos experimentalmente o impacto de prováveis marcadores de autoria, pretendemos cumprir tais objetivos e contribuir assim para o desenvolvimento dos estudos de atribuição de autoria em Linguística Forense no Português Europeu.

No próximo capítulo, apresentamos uma breve contextualização histórica da Linguística Forense e expomos o estado da arte no contexto português. Depois, no capítulo 3, elaboramos algumas considerações sobre a apropriação dos termos "idioleto", "estilo" e "estilo idioletal" pelos estudos de atribuição de autoria e levantamos algumas questões em torno da seleção de marcadores de autoria. No capítulo 4 explicamos a teoria de base das Máquinas de Vetores de Suporte e a sua adequação aos estudos de atribuição de autoria em Linguística Forense. No capítulo 5 descrevemos a experiência desenvolvida no âmbito desta dissertação, cujas notas conclusivas surgem depois, já no capítulo 6.

2 – Linguística Forense – História e Enquadramento

A Linguística Forense é uma disciplina em que se utilizam conhecimentos da Linguística para a peritagem sobre o uso da língua em contextos de criação, observação e aplicação da lei. Consequentemente, a investigação em Linguística Forense é o resultado da articulação entre várias áreas de conhecimento,[1] se bem que com destaque para estas duas: o Direito e a Linguística. Pode ter uma grande variedade de aplicações, incidindo essencialmente sobre a linguagem escrita da Lei (por exemplo, na compreensão ou interpretação da Lei), sobre a linguagem dos processos legais (como a que se pode encontrar nas atas dos tribunais ou nas transcrições dos interrogatórios policiais) e sobre enunciados linguísticos que funcionem como prova em contexto judicial (caso das questões de atribuição de autoria a enunciados da escrita ou da fala, da deteção de plágio ou da disputa de direitos de autor)[2].

Embora os termos "Linguagem e Direito" e "Linguística Forense" inicialmente se referissem a áreas de investigação com incidências distintas – a Linguagem e Direito sobre questões de elaboração e interpretação da Lei e a Linguística Forense sobre a análise linguística de provas judiciais –, o termo Linguística Forense tem triunfado sobre o anterior, ganhando, como explica Gibbons (2003:12), um sentido cada vez mais lato: "The term 'Forensic Linguistics' can be used narrowly to refer only to the issue of language evidence. However it is becoming accepted as a cover term for language and the law issues".

A história desta área de investigação começou há algumas décadas, na sequência de uma acumulação de erros judiciais que se cometeram por ausência de peritagens linguísticas, uma vez que também não era evidente que seria necessário auscultar a opinião de linguistas no âmbito do exercício da Lei e da aplicação da Justiça.

Foi no Reino Unido e nos Estados Unidos que surgiram os primeiros estudos significativos na área, e em 1968, no Reino Unido, recorria-se pela primeira vez ao termo "Linguística Forense" num artigo de Jan Svartvik: *The Evans Statements: a Case for Forensic Linguistics*. A

[1] Cf. Coulthard and Johnson (2007:6) *"Early forensic linguistic research originated in a wide range of disciplines(...). Research since 1990 has continued to come from all these disciplines, making forensic linguistics a multi- and cross-disciplinary field."*

[2] Conforme consulta do *site* da IAFL – International Association of Forensic Linguists (About Us), http://www.iafl.org/forensic.php (consulta em 27/02/2015). Cf. Coulthard and Johnson (2007:5)

análise linguística elaborada por Svartvik (1968) permitiu concluir que o grupo de declarações não-condenatórias de Timothy Evans era discrepante, estilisticamente, em relação ao conjunto das declarações que serviram para o incriminar. Timothy Evans fora executado, mas foi ilibado postumamente. A sua inocência já tinha sido confirmada porque John Christie, o verdadeiro assassino, confessara depois de indiciado e condenado por um conjunto de assassinatos em série.

A injusta condenação à morte de Timothy Evans contribuiu para o debate público que culminou, em 1965, na abolição da pena capital no Reino Unido. Este e outros erros na aplicação da Justiça começaram a despertar a comunidade jurídica para a necessidade de recorrer a pareceres elaborados por peritos em linguística. Tratava-se de exigir pareceres que, por um lado, acrescentassem uma análise bem fundamentada, baseada em dados abalizados e não derivada do senso comum, mas também que, por outro lado, fossem elaborados por especialistas externos que pudessem investigar questões de legitimidade probatória, como no caso das declarações obtidas por coação. Isto contribuiu para a afirmação desta área de investigação e para, progressivamente, começarem a desenvolver-se mais e melhores metodologias para a análise linguística com aplicação judicial.

Apesar de o contributo do linguista como testemunha pericial em tribunal ainda ser um contributo limitado, todo este processo alimenta, ao mesmo tempo que enriquece, as áreas de estudo em Linguagem e Direito e Linguística Forense, as quais, nos últimos anos, têm crescido exponencialmente, a par de um aumento de contribuições especializadas de índole académica. É este o caso dos manuais dedicados à Linguística Forense como Coulthard e Johnson (2010), Gibbons e Turell (2008), Coulthard e Johnson (2007), Olsson (2004) e Gibbons (2003), bem como o livro de McMenamin (2002), numa perspetiva mais centrada na estilística forense. Destacam-se também duas revistas científicas subordinadas ao tema: *The International Journal of Speech, Language and the Law* (International Association of Forensic Linguists) e *Language and Law/Linguagem e Direito* (Faculdade de Letras da Universidade do Porto e Universidade Federal de Santa Catarina).

2.1 – A Linguística Forense no contexto português

Em Portugal, a Linguística Forense começou a dar os primeiros passos há algumas décadas, inicialmente com incidência na Fonética Forense, com os primeiros trabalhos a serem desenvolvidos em Fonética Acústica por Maria Raquel Delgado-Martins. O trabalho de peritagem nesta área tem sido continuado por Fernando Martins que, inclusivamente, criou o Núcleo de Investigação em Fonética Forense (NIFF), grupo que visa estabelecer a ponte entre a investigação sobre o tema e a sua aplicação ao contexto judicial. Fazem parte deste núcleo Fernando Martins, Celeste Rodrigues, Fernando Brissos (do Centro de Linguística da Universidade de Lisboa) e Deolinda Simões (perita em Direito e Ciências Forenses e técnica superior da Administração Tributária do Ministério das Finanças). Alguns artigos referentes ao tema foram publicados recentemente, com destaque para Martins *et al.* (2014), no qual se destaca o isolamento de um traço fonético particular, de natureza não forjável, que foi testado com sucesso como método de identificação do falante. Gillier (2011) apresentou também um contributo para esta área, instrumentalizando a fonética acústica para analisar o efeito que alguns disfarces da voz exercem na sua frequência fundamental.

Mais recentemente, a Linguística Forense no contexto português enriqueceu-se com uma contribuição académica significativa, a revista *Linguagem e Direito*, editada pela Faculdade de Letras da Universidade do Porto e pela Universidade Federal de Santa Catarina, com Malcolm Coulthard e Rui Sousa-Silva como editores. Esta publicação bianual disponibiliza artigos em inglês e em português e "tem como objetivo impulsionar a disseminação da pesquisa nos domínios da Linguística Forense / Linguagem e Direito e, ao mesmo tempo, contribuir para o exercício da prática na área, pela publicação de artigos sobre o estado da arte de questões teóricas e de ferramentas metodológicas aplicáveis a esse campo interdisciplinar."[3]

Além desta publicação, contemplando também a área da "Linguagem e Direito" (ou "Linguagem da Lei"), Rodrigues (2005) considerou um corpus de discurso em contexto de

[3] Descrição dos objetivos da publicação, conforme consulta do site www.linguisticaforense.pt

tribunal, numa perspetiva que privilegiou a análise do discurso oral. Nesta obra podem observar-se considerações relevantes sobre a especificidade da linguagem jurídica e sobre as dificuldades de comunicação que advêm do seu uso. A autora também se debruçou sobre a melhor forma de a linguística poder participar tanto na legislação, como na sala de audiências, sobretudo em questões que impliquem assegurar os direitos dos cidadãos com língua materna distinta da que é usada no exercício da Lei, os quais, por isso mesmo, podem ter necessidade de um intérprete.

No que respeita à deteção de plágio, em Sousa-Silva (2013) são detalhadamente exploradas algumas formas de plágio e, mais concretamente, o que pode ser dito e feito quanto à intencionalidade e não intencionalidade do plágio, bem como à imputação legal de quem o comete. Considerando a atribuição de autoria a textos escritos de caráter não-literário, a dissertação de Sousa-Silva foi pioneira para o português europeu. Por incluir uma análise apoiada em ferramentas da linguística computacional e em métodos estatísticos, que permitem detetar plágio sem ser necessário haver uma confrontação textual com correspondência *verbatim,* Sousa-Silva (2013) acrescentou uma vertente mais quantitativa aos critérios de atribuição de autoria, contribuindo para uma crescente credibilização do processo de deteção de plágio no contexto português. O autor conta também com outras contribuições significativas que alargam os processos de atribuição de autoria a outras plataformas, nomeadamente algumas plataformas *online* bastante prolíferas na produção de textos escritos de pequena dimensão, como é o caso do *Twitter* (cf. Sousa-Silva *et. al.* (2011)), tópico especialmente relevante dado que a dimensão deste tipo de texto se aproxima muito daquela que se pode esperar em contexto judicial real. Além de fazer uma análise computacional, Sousa-Silva testou o grau de sucesso da atribuição de autoria recorrendo a diferentes traços estilísticos, e os resultados obtidos foram bastante expressivos, apesar dos constrangimentos estruturais da plataforma do Twitter, que apenas permite a produção de textos com um máximo de 140 caracteres.

Ainda sobre a atribuição de autoria em textos escritos, podemos encontrar outro contributo significativo no trabalho desenvolvido em Marquilhas e Cardoso (2011), o qual apresenta os resultados de um estudo de caso que conjuga uma análise quantitativa com uma análise qualitativa num caso de atribuição de autoria a uma crónica caluniosa que fora

previamente atribuída a uma conhecida jornalista. Através da realização de uma experiência com dados reais, as autoras elaboraram um *corpus* com textos de caráter cronístico dos dois autores suspeitos de serem responsáveis pela elaboração do texto questionado (texto Q). Recorrendo a programas de estatística lexical e isolando variáveis textuais já usadas em outras áreas (como no estudo discursivo de corpora, na psicologia social e na estilística forense), o corpus reunido foi submetido a um estudo contrastivo de "originalidade" (*keyness*). Foi ainda elaborada uma análise qualitativa com base em pontuação e sintaxe. As suas conclusões adicionaram um fator importante a considerar em experiências futuras sobre a atribuição de autoria: o tipo de texto considerado no artigo, i.e., a crónica, tem intrinsecamente um estilo fortemente tipificado e isso aproxima também o estilo dos seus autores.

Apesar de a análise quantitativa ser uma mais-valia na investigação em atribuição de autoria a textos escritos, uma análise qualitativa não pode ser descartada. As conclusões em linguística forense devem ser expressas numa gradação de probabilidade e, por isso, uma abordagem centrada num conjunto fechado de variáveis não é a mais acertada. Conforme Marquilhas e Cardoso (2011) afirmam, com a internet cresceu a facilidade de produção e divulgação de informação. Consequentemente, cresceu também a possibilidade de fraude envolvendo textos escritos. Neste sentido, a investigação em linguística forense precisa de cada vez mais contribuições, e contribuições que sejam interdisciplinares.

"A linguística já desenvolveu uma série de disciplinas que podem apoiar a investigação destas fraudes. Trata-se agora de articular os axiomas de cada uma delas e de problematizar a forma como eles se complementam no contexto deste desafio. As disciplinas em causa são sobretudo a análise do discurso, a pragmática, a sintaxe, a crítica textual, a linguística histórica, a sociolinguística e a linguística de corpus."

Marquilhas e Cardoso (2011: 418)

O trabalho que esta dissertação pretende desenvolver inclui-se nesta 'articulação de axiomas'. Conjuga uma análise quantitativa com a análise qualitativa e ocupa-se da atribuição de autoria a textos escritos, textos esses que poderiam funcionar como provas em contexto judicial.

3 – Idioleto, estilo e estilo idioletal

"The community is prior to the individual. Or to put it another way, the language of individuals cannot be understood without the knowledge of the community of which they are members" Labov (2006 (1966):5).

Partindo desta perspetiva laboviana de idioleto enquanto "language of the individual", uma variedade muito carregada de informação coletiva, Turell (2010) apresenta uma breve discussão sobre a adequação do termo idioleto aos contextos de atribuição de autoria em linguística forense [4]. No seu artigo, Teresa Turell acaba por defender que há vantagem em falar antes em "estilo idioletal". O estilo idioletal demarca-se do conceito de idioleto por ter o seu foco não no sistema linguístico do indivíduo, mas no uso que o indivíduo faz do sistema linguístico que partilha com a sua comunidade. Defende-se também que o estilo idioletal terá maior variação interautores do que intra-autor, mesmo considerando diferentes tipos textuais, com exclusão apenas das expressões formulares e do vocabulário de textos especializados. Para Turell trata-se de um uso que envolve algum arbítrio individual, extensivo a um "conjunto de opções": "Thus, in the context of forensic text comparison, 'idiolectal style' could be defined as the set of options that writers take from the linguistic repertoire available to them as users of a specific language".

A matização envolvida nesta proposta de Turell explica-se porque classificar e delimitar estritamente a variedade linguística de um indivíduo — o seu idioleto — seria o equivalente a assumir que se poderia fixar um perfil textual exclusivo para cada pessoa que produzisse um enunciado escrito. Em vez disso, o que a autora propõe que se apure são os elementos textuais que podem ser isolados para um indivíduo, sendo que a validade dos resultados varia na razão direta da quantidade de dados disponível (Turell (2010:217)). E é preciso ter sempre presente, ainda, que se trata de dados voláteis, facilmente sujeitos a mudanças devido ao inevitável e constante contacto dos falantes com a sua comunidade linguística. Outros autores, com outra terminologia mas a mesma atitude, falam em "padrões de

[4] Para uma análise mais profunda sobre o conceito de idioleto na história da linguística moderna, cf. De Beaugrande (1998) "Language and Society: the real and the ideal in linguistics, sociolinguistics and corpus linguistics." Em *Journal of Sociolingistics (3)1: 128-139.*

elementos distintivos" (Johnson e Wright (2014:39)) ou então em "consistência". Conforme Grant (2010: 509): “Practical authorship analysis may depend less on a strong theory of idiolect than on the simple detection of consistency and the determination of distinctiveness”.

Nesta perspetiva, englobamos na nossa análise forense o conceito de **estilo**. Algumas das aceções linguísticas que este termo encerra podem ser encontradas em Coutinho (2002): “qualquer produção linguística implica escolhas (mais ou menos conscientes), que correspondem a um trabalho de formulação a que, em última análise, se poderá chamar estilo”. Já na perspetiva da linguística forense, McMennamin (2002) resume: “Style in writing results from the recurrent choices that the writer makes.” Entende-se portanto que o estilo está intrinsecamente relacionado com as **escolhas** que cada indivíduo tendencialmente faz na produção dos seus enunciados, considerando que as escolhas incidem sobre as opções disponíveis para a sua própria língua.

Admitimos também que em contextos forenses, tipicamente, não existem dados suficientes para fazer uma caracterização global do estilo idioletal de cada falante, mas mesmo assim a identificação do autor de um enunciado pode ser tentada recorrendo ao levantamento de um conjunto de marcadores de autoria que podem ser obtidos nas produções textuais, marcadores que permitirão tentar o traçado de um estilo idioletal.

3.1 – Marcadores de autoria

O problema típico de atribuição de autoria em contextos forenses implica identificar o autor a partir de um conjunto limitado de textos e de suspeitos (Stamatatos (2009:2), Luyckx and Daelemans (2008:513); Coulthard (2006:2)). A investigação que tem sido feita neste sentido tem articulado diferentes marcadores de autoria com diferentes metodologias. Apesar das diferenças de contextos, é possível identificar alguns dos marcadores mais bem sucedidos num grande número de experiências (Grant and Baker (2001:68), Stamatatos *et al.* (2001), Diederich *et al.* (2003)).

A cosseleção recorrente de opções linguísticas contribui para a delineação de um perfil ou estilo idioletal. No caso das escolhas lexicais de cada indivíduo, por exemplo, Coulthard (2006:1) sublinha: "Thus, whereas in principle any speaker/writer can use any word at any time, in fact they tend to make typical and individuating co-selections of preferred words." O propósito do investigador em linguística forense, enquanto testemunha pericial[5], é o de conseguir aproximar-se desse perfil, começando por identificar nos textos questionados quais as escolhas linguísticas do indivíduo que, por constituírem variações à norma e lhe serem particulares, podem ser classificadas como marcadores de autoria. A escolha de determinadas variáveis linguísticas em detrimento de outras pode ajudar a identificar algumas informações extralinguísticas sobre o autor de um enunciado questionado, conforme defende Turell (2010: 212):

> "Forensic linguists work with the assumption that linguistic production of individual speakers and writers can sometimes reveal information about an individual's age, gender, occupation, education, religion and political background. It can also provide clues to the determination of an individual's geographical origin, ethnicity or race."

Estes dados que podem emergir nas produções textuais podem ser captados devido ao desenvolvimento dos estudos linguísticos, por exemplo, no âmbito da sociolinguística, da dialetologia e da aquisição de L2, i.e. da aquisição de uma segunda língua além da língua materna, (Turell (2010:220)). Porém, surgem alguns obstáculos nestes processos quando se passa aos casos forenses reais. Aqui as provas escritas são muitas vezes escassas e breves, não havendo oportunidade para o afloramento deste tipo de variação. E mesmo quando isso acontece, i.e., quando se detetam itens marcados em termos de estilo, é necessário usar de bastante moderação ao propor a identificação do autor de um texto no caso dos enunciados de valor probatório, uma vez que as consequências de uma peritagem falaciosa podem refletir-se em penas injustas: "While it is possibly true that mistakes made by authorship analysts in the field of literature could lead to red faces and bad press at worst, the same cannot be said of the forensic context, where mistakes could lead to imprisonment or even

[5] Considerámos "testemunha pericial" conforme a descrição de "expert witness" em Coulthard (2010:478).

execution in certain countries. The importance of extreme caution before arriving at conclusions can therefore not be overemphasised" (Kotzé (2010:186)).

Por outro lado, a escolha dos marcadores de autoria deve ter em consideração factos que possam ser um obstáculo a identificações adequadas, como por exemplo a introdução de itens de disfarce, muito dependentes do talento metalinguístico de cada um. Embora a linguagem seja uma capacidade inata do ser humano, a frequência escolar obrigatória inculca nos indivíduos um conjunto normalizado de regras, especialmente relativas à enunciação da língua escrita (Castro (2006)); além disso, imprime-lhes consciência metalinguística, que se torna proporcional ao grau de escolaridade, ou nível de literacia. Conforme defendido por McCombe (2002:6), este facto traz desafios para a atribuição de autoria, uma vez que, logicamente, um autor com maior domínio da língua terá uma maior capacidade de introduzir disfarce nas suas produções textuais. É por isso que os itens considerados como marcadores do estilo idioletal são os mais dificilmente forjáveis, como é o caso de algumas variáveis que envolvem estrutura sintática e o uso de determinados itens morfossintáticos ((Chaski (1997:19); McCombe (2002:5)).

Nos últimos anos, a investigação em atribuição de autoria tem enveredado crescentemente para o investimento nas análises quantitativas, a par das mais tradicionais análises qualitativas. Pretende-se, idealmente, obter métodos que recorram a marcadores de autoria discriminantes, métodos que possam ser replicados e que, por conseguinte, aumentem a fiabilidade dos resultados. O conjunto destas duas análises permitirá definir mais abalizadamente o estilo do autor e, finalmente, contribuir para o objetivo último da peritagem linguística em contexto forense, que consiste em dar uma resposta confiável à pergunta "Quem escreveu este texto?".

No próximo capítulo analisaremos alguma fundamentação teórica de base para a análise quantitativa da experiência realizada nesta dissertação. De seguida, apresentaremos a metodologia, a amostra, e os testes experimentais efetuados na análise quantitativa, antes de procedermos à análise qualitativa.

4 – Máquinas de Vetores de Suporte em Linguística Forense

Conforme Coulthard e Johnson (2007), uma das primeiras abordagens de base mais estatística em atribuição de autoria remonta a 1851. Augustus De Morgan, numa tentativa de atribuir autoria a duas epístolas bíblicas de São Paulo, sugeria comparar-se a média de letras por palavra de dois livros bíblicos, sendo que a proximidade de resultados significaria o mesmo autor para ambos os textos. Posteriormente, Mosteller e Wallace (1964) e Kenny (1982) tentaram também uma análise puramente estatística para questões de atribuição de autoria. No entanto, conforme Olsson (2008:19) afirma, existem necessariamente lacunas nos métodos puramente estatísticos; sobretudo, não podem ser aplicados sem conhecimento do funcionamento da língua, uma vez que existe a necessidade de garantir que as variáveis consideradas nestes testes são marcadores de autoria relevantes.

Nos últimos anos, muitos dos trabalhos de investigação com resultados mais pertinentes têm recorrido a testes estatísticos conjugados com métodos computacionais. Conforme Koppel *et al.* (2009), o problema típico de atribuição de autoria, o qual compreende um conjunto definido de autores possíveis para a atribuição de um texto questionado, é, em última análise, um problema de categorização textual.

A identificação de padrões textuais com recurso a modelos algorítmicos desenvolvidos para processamento de linguagem natural, mais especificamente *text-mining*, está amplamente testada e estabelecida. É aqui que se enquadram as Máquinas de Vetores de Suporte (SVMs, do inglês *Support Vector Machines*), usadas para reconhecimento de padrões em imagens, para bioinformática, e também para categorização de textos (Lorena e Carvalho (2003)).

Os conceitos de base das Máquinas de Vetores de Suporte foram desenvolvidos por Vapnik (1995) e trabalham conforme a descrição em De Vel *et al.* (2001): "The SVMs' concept is based on the idea of structural risk minimisation which minimises the generalisation error (i.e. true error on unseen examples) (...) The use of a structural risk minimisation performance measure is in contrast with the empirical risk minimisation approach used by conventional classifiers. Conventional classifiers attempt to minimise the

training set error which does not necessarily achieve a minimum generalisation error. Therefore, SVMs have theoretically a greater ability to generalise."

A abordagem das Máquinas de Vetores de Suporte supõe um processo prévio de aprendizagem automática (*Machine Learning*) através de indução de um classificador automático, de forma a que este possa fazer uma identificação binária entre padrões, ou seja, classificar entre apenas duas opções possíveis, segundo as amostras dadas para treino. Utilizando uma Máquina de Vetores de Suporte, é possível classificar instâncias a partir de quaisquer elementos dos domínios em que o classificador da SVM foi treinado. Ao conseguir identificar uma margem máxima de separação entre os pontos (Fig.1) de dois conjuntos de dados, desenha-se uma linha de fronteira, um hiperplano, de forma a que seja possível atribuir uma de duas classes a qualquer novo ponto que seja processado pela Máquina de Vetores de Suporte. Seguindo este método, a performance obtida é superior, mesmo considerando um grande número de elementos distintivos que atuem como coordenadas destes pontos, uma vez que o foco da classificação não está no entrecruzamento das classes, mas sim no estabelecimento de uma margem máxima de separação entre os planos que definem as classes.

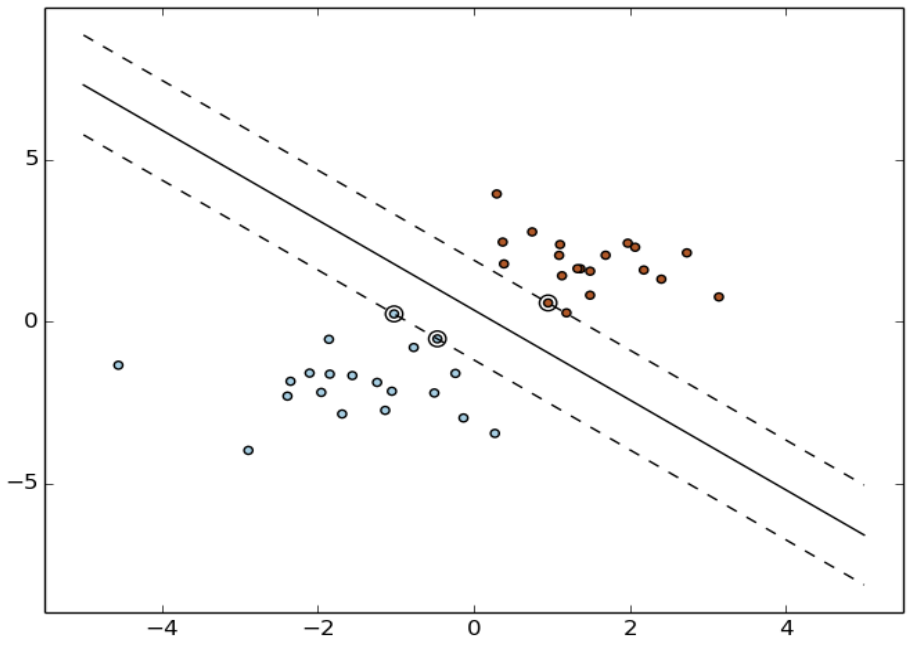

Fig. 1 – Exemplo de margens máximas e hiperplano a separar dois conjuntos de dados com recurso a uma Máquina de Vetores de Suporte. (Fonte: scikit-learn.org)

No processamento textual em atribuição de autoria, as Máquinas de Vetores de Suporte consideram cada texto como um ponto (ou vetor) cujas coordenadas correspondem a dimensões. As dimensões, por sua vez, são correspondentes a um número variável de elementos distintivos e contabilizáveis, que são determinados pelo usuário e calculados para cada texto. Alguns exemplos desses elementos poderão ser o número de ocorrências de uma palavra, a dimensão média de cada frase ou o número de *tokens* de um texto.

Por serem múltiplos os fatores que se podem considerar para a atribuição de autoria textual com recurso a Máquinas de Vetores de Suporte, há notícia de várias experiências em que se testaram diferentes elementos estilísticos enquanto marcadores de autoria (Grant e Baker (2001:66)), com um diverso número de autores e com *corpora* de dimensões distintas para cada autor (Fissette (2010:7)).

Em contextos forenses reais, o número de marcadores, o número de autores e a dimensão do texto escrito são variáveis que se manifestam de forma imprevisível, se bem que o universo de suspeitos seja normalmente muito limitado (Luyckx e Daelemans(2008:1), Koppel *et. al.* (2009:2, 3))[6]. Relativamente aos elementos estilísticos, já se testou, e com boas taxas de sucesso, um conjunto significativo de marcadores de autoria. Porém, segundo Grant e Baker (2001:69), existem alguns perigos associados ao processo de escolha dos marcadores mais adequados: por um lado, o sucesso de uns marcadores de autoria num conjunto de textos não garante que estes marcadores sejam igualmente bem-sucedidos noutra amostra textual; por outro lado, poder-se-á assumir precipitadamente uma maior fiabilidade de determinados marcadores comparativamente a outros.

Tem-se verificado, contudo, que nos casos que incluem a utilização das Máquinas de Vetores de Suporte, não é preciso travar a escolha de marcadores de autoria, dado que uma maior quantidade de traços estilísticos reunidos parece contribuir para uma maior taxa de sucesso (Sousa-Silva *et al.* (2011), Hirst e Feiguina (2007), De Vel *et al.* (2001)). Adicionalmente, a utilização de Máquinas de Vetores de Suporte na investigação da atribuição de autoria parece bem talhada para a aplicação a casos judiciais, já que permite

[6] Cf. Koppel et. al. (2009) para uma consideração sobre outros cenários possíveis relativamente à dimensão da amostra.

lidar com textos de dimensão reduzida, quer no caso dos textos de autoria questionada, quer no caso das amostras textuais disponíveis para o *corpus* de treino dos autores questionados. Caberá ao investigador determinar estas variáveis e selecionar um teste que seja replicável, distintivo, confiável, e que possa servir um grande número de casos.

Em De Vel *et al.* (2001) foram considerados textos de mensagens de correio eletrónico de três autores diferentes, com cerca de 12.000 palavras para cada autor. A experiência contemplou mensagens aglomeradas sob o mesmo assunto e mensagens multitópico. Nesta experiência, as mensagens testadas tinham cerca de 156 palavras, e, para marcadores de autoria, foi selecionado para um primeiro teste um conjunto de características que descreviam a estrutura de cada email, consideradas por isso como características estruturais. Num segundo teste, foram usados marcadores estilísticos tais como o número de palavras gramaticais, a média de tamanho de frase, o número de caracteres de pontuação, etc. Estes traços foram analisados com recurso a uma Máquina de Vetores de Suporte e os resultados obtidos foram mais bem sucedidos no teste com marcadores estilísticos do que no teste com elementos estruturais, mas os resultados da reunião dos elementos de ambos os testes superaram os dois anteriores.

Em Sousa-Silva *et al.* (2011) os autores testaram a atribuição de autoria a porções textuais de editoriais de um jornal português com recurso a uma Máquina de Vetores de Suporte, sendo que as porções textuais que foram usadas como textos questionados eram frases soltas. Os marcadores de autoria selecionados para o teste na SVM foram divididos em subgrupos – um subgrupo com marcadores de autoria baseados no conteúdo lexical e outro subgrupo com marcadores de autoria de baseados em elementos estruturais. O seu desempenho foi testado, quer para os subgrupos, quer para o conjunto geral, e concluiu-se que os traços estruturais, baseados em etiquetas morfológicas, pontuação e dimensão de frase ou palavra, contribuíam de forma mais significativa para a atribuição de autoria ao nível da frase, superando os resultados obtidos pelos traços baseados em conteúdo lexical. O teste que uniu os dois subgrupos de traços foi o teste com melhor desempenho, comprovando a ideia de que uma maior quantidade de traços contribui para um aumento do potencial discriminatório da Máquina de Vetores de Suporte.

Em Hirst e Feiguina (2007) foi desenvolvida uma experiência que pretendia identificar a autoria de porções textuais de 1.000, 500 e 200 palavras que poderiam ter sido escritas ora por Anne Brontë, ora por Charlotte Brontë. Estas autoras oferecem reconhecidas dificuldades de distinção quando são usados apenas os métodos tradicionais. Os conjuntos originais para treino eram de grande dimensão, com cerca de 250 mil palavras para cada autora. Na análise quantitativa desenvolvida pelos investigadores, usaram-se como traços discriminatórios bigramas de etiquetas sintáticas, conseguidos através de uma operação de *parsing* parcial dos textos, tratando-se posteriormente como unidades os sucessivos fragmentos obtidos. Adicionalmente, escolheram-se marcadores de autoria tais como a frequência de etiquetas morfossintáticas, o comprimento médio de palavra e o comprimento médio de frase. As contagens relativas das suas frequências atuaram como coordenadas nos vetores usados para o classificador da Máquina de Vetores de Suporte. Os resultados comprovaram que, mesmo em porções textuais de pouco mais de 200 palavras, o somatório dos marcadores de autoria apresentava um desempenho superior ao de qualquer dos conjuntos de marcadores tomados de forma individual, como acontecera nas experiências anteriores.

Nesta dissertação pretende-se, usando como fundamentação teórica essencial os trabalhos de investigação citados, testar a aplicabilidade de uma análise quantitativa computacional recorrendo a Máquinas de Vetores de Suporte para atribuição de autoria. Numa segunda fase, pretende-se simular um caso forense de atribuição de autoria e resolvê--lo combinando métodos qualitativos com métodos quantitativos.

5 – Experiência

5.1 – Metodologia

Escolhemos uma Máquina de Vetores de Suporte para o processamento automático dos dados na experiência desta dissertação devido aos argumentos acima apresentados: com efeito, trata-se de um classificador que apresenta boa capacidade de generalização, além de se basear numa teoria estatística e matemática bem definida (Smola *et al.* (1999) *apud.* Lorena e Carvalho (2007)). Optámos pela Máquina de Vetores de Suporte da aplicação *Scikit-learn* (Pedregosa *et al.* (2011)), que está disponível gratuitamente, e aplicámo-la a um corpus que passaremos a apresentar.

O investigador que procure especializar-se em atribuição de autoria na área dos estudos forenses encontra alguns constrangimentos legais relativamente ao acesso aos materiais dos casos judiciários reais. Por um lado, muitos dos textos que poderiam servir como amostra têm valor probatório e, por isso, não poderiam ser facultados nem tornados públicos em resultado da investigação. Por outro lado, os resultados conseguidos, não sendo vinculativos, podem interferir nos julgamentos por estabelecerem apreciações quanto aos seus hipotéticos autores. Adicionalmente, na realização de uma experiência científica pressupõe-se isolar, delimitando e descrevendo, a eventual interferência de fatores externos que possam moldar os resultados, o que não se coaduna com a variabilidade dos dados em contexto legal.

Reconhecendo a possível interferência de fatores como a idade, o género e as habilitações académicas nos testes de atribuição de autoria, Carole Chaski compilou um conjunto de textos produzidos por informantes de um perfil sociológico controlado para os testes de atribuição de autoria que publicou entre 1997 e 2006, o que lhe permitiu excluir géneros textuais muito díspares e aumentar a potencial significação dos marcadores testados.

No mesmo sentido, compilámos aqui um *corpus* textual que permitisse o controlo de fatores externos. Selecionámos um grupo de informantes de perfil sociolinguístico

controlado, sujeitos esses que aceitaram colaborar através da elaboração das amostras textuais.

Para conseguirmos controlar influências variáveis e obter uma interpretação mais clara dos resultados da experiência desta dissertação, dividimos o processo experimental em dois testes distintos: 1) um teste que verificasse a eficiência do método de atribuição de autoria computacional que selecionámos, i.e., a Máquina de Vetores de Suporte, sobre o *corpus* recolhido; 2) um teste com proximidade ao contexto forense real, em que, para um texto questionado, fosse verificável a probabilidade de o classificador acertar no seu real autor.

5.2 – Constituição da Amostra

Com os testes que pretendíamos realizar, queríamos verificar se seria possível usar certos marcadores de autoria para identificar o autor de um texto; queríamos também verificar o grau de influência de certas variáveis, como a formação curricular e o género textual, na produção de textos.

Reunimos um conjunto com doze informantes por considerarmos que dessa forma obteríamos dados textuais em dimensão razoável para avaliar a influência das diferenças de género (seis homens e seis mulheres) e das diferenças ao nível da formação curricular (seis informantes eram licenciados em Ciências e seis em Letras). Para o efeito, reunimos uma amostragem não probabilística por escolha racional, selecionando os indivíduos que obedecessem aos critérios estipulados de acordo com o seguinte perfil:

- 12 informantes
    - 6 informantes licenciados na Faculdade de Ciências da Universidade de Lisboa

    (3 homens e 3 mulheres)
    - 6 informantes licenciados na Faculdade de Letras da Universidade de Lisboa

    (3 homens e 3 mulheres)
- Idades entre os 20 e os 35 anos de idade
- Naturais da área metropolitana de Lisboa
- Habilitações académicas: licenciatura concluída e ainda ligados à investigação académica (estudantes ou bolseiros de investigação).

Exigiu-se que o percurso académico fosse homogéneo, i.e., que cada informante tivesse frequentado o ensino secundário numa área relacionada com a área em que prosseguiu os estudos na formação universitária. Ao conjunto de informantes descrito foi pedida a redação anónima de quatro textos: uma carta de reclamação, uma carta de ameaça, uma carta de extorsão e uma carta de agradecimento. Os quatro textos teriam de ser redigidos em computador, no mesmo dia, em documentos individuais com um mínimo de 300 palavras cada, usando o mesmo editor de texto e com o mesmo tipo e corpo de letra.

A recolha foi ajustada conforme a disponibilidade de cada autor, num intervalo de aproximadamente 2 meses, entre Maio e Julho de 2015.

5.3 – Amostra

Id. de informante[7]	n.º de palavras na c. reclamação	n.º de palavras na c. ameaça	n.º de palavras na c. extorsão	n.º de palavras na c. agradecimento	Total de palavras
FC_F23_DO	327	305	352	314	1298
FC_F24_JV	301	308	302	284	1195
FC_F29_SA	300	302	318	337	1257
FC_M23_FA	306	317	310	374	1307
FC_M26_MG	318	306	299	299	1222
FC_M28_AN	423	324	511	345	1603
FL_F29_DM	314	309	329	313	1265
FL_F31_AC	315	315	358	330	1318
FL_F31_NB	279	314	290	290	1173
FL_M26_JC	308	316	410	401	1435
FL_M32_PO	303	306	300	304	1213
FL_M34_BH	297	309	308	301	1215
Média aritmética simples	316	311	341	324	1292
Total	3791	3731	4087	3892	15501

Tabela 1 – Número de palavras dos textos que compõe a amostra

5.4 – Análise Quantitativa

A etiquetação sintática e morfossintática dos textos, bem como a programação da Máquina de Vetores de Suporte, contou com a colaboração do Doutor João Silva, do grupo "NLX – Natural Language and Speech Group" da Faculdade de Ciências da Universidade de Lisboa.

A constituição do *corpus* pretendeu aproximar-se de um cenário verosímil em termos de representatividade textual, uma vez que os textos que constituem as provas textuais são muitas vezes de tamanho reduzido. Nesse sentido, optámos por considerar um pequeno conjunto de textos curtos (com cerca de 300 palavras cada) para treino do classificador. Os textos foram etiquetados com informação morfossintática (Fig.2) e sintática (Fig.3), de

[7] Legenda da identificação dos informantes: Instituição de Origem_Género/Idade_Iniciais de Identificação

acordo com o sistema de etiquetas do Corpus Internacional do Português – CINTIL (Barreto *et. al*. (2006)), desenvolvido em colaboração pelo grupo NLX – Natural Language and Speech Group da Universidade de Lisboa e pelo CLUL – Centro de Linguística da Universidade de Lisboa. O seu POS-tagger, LX-suite, tem uma taxa de acerto de 97% (Branco e Silva (2006)) e o seu *parser* de constituência*,* LX-parser, atinge um desempenho de 88% F_1 ((Silva *et al.* (2010)).

Nos testes realizados, selecionámos marcadores de autoria contabilizáveis de acordo com o que vem sugerido na bibliografia anteriormente indicada:

- Bigramas e trigramas de etiquetas POS (*part-of-speech)*
- Bigramas e trigramas de categorias sintáticas
- Contagens de itens de pontuação
- Comprimento médio de frase
- Contagens de itens lexicais

O *Scikit-learn* foi programado de forma a que o texto de input fosse processado nos seguintes módulos sequenciais:

- "CountVectorizer", para transformar texto num vetor de contagens absolutas;
- "TfidfTransformer", para converter um vetor de contagens absolutas num vetor de medidas de relevância;
- "SGDClassifier", que atua como o classificador propriamente dito da Máquina de Vetores de Suporte.

Para o Teste I separou-se o conjunto das 12 cartas de ameaça (*corpus* de teste) do conjunto das restantes cartas redigidas pelos doze autores (*corpus* de treino), e verificou-se a possibilidade de atribuir as variáveis "formação curricular, "género" e "autoria" corretas às cartas de ameaça dos informantes. No caso do Teste II, todo o conjunto das 48 cartas foi usado como *corpus* de treino, enquanto uma carta adicional funcionava como texto questionado, ou "TextoQ" (texto de teste). Para ambos os testes (Teste I e Teste II) criou-se um objeto Python "CountVectorizer" que, ao ser aplicado a um texto, o converte num vetor

de contagens. Posteriormente, aplicou-se-lhe a transformação "TfidfTransformer", que permite reduzir o impacto de palavras muito frequentes que não tenham valor discriminatório significativo, à medida que aumenta o valor discriminatório de palavras menos frequentes. A métrica tf-idf (do inglês, *term frequency – inverse document frequency*) é bastante usada em operações de *text-mining* pois permite salientar a importância de uma palavra num documento em relação a outro conjunto de documentos, recalibrando o seu valor em relação à sua preponderância nos restantes documentos.

Bigramas e trigramas de Categorias Morfossintáticas (POS) e Sintáticas

Após o processo de etiquetagem dos textos com informação morfossintática POS (*part-of-speech*) os textos foram esvaziados do seu conteúdo lexical de forma a que ficassem apenas as etiquetas POS no lugar das palavras originais.

```
Espero que esta carta seja o ponto final de uma história incómoda, que em nada dignifica
os serviços que prestam.

Espero/ESPERAR/V#pi-1s que/QUE/CJ esta/ESTA/DEM#fs carta/CARTA/CN#fs seja/SER/V#pc-3s
o/O/DA#ms ponto/PONTO/CN#ms final/FINAL/ADJ#ms de/DE/PREP uma/UMA/UM#fs
história/HISTÓRIA/CN#fs incómoda/INCÓMODO/ADJ#fs ,*//,/PNT que/QUE/REL em/EM/PREP
nada/NADA/IND#ms dignifica/DIGNIFICAR/V#pi-3s os/OS/DA#mp serviços/SERVIÇO/CN#mp
que/QUE/REL prestam/PRESTAR/V#pi-3p .*//./PNT

V CJ DEM CN V DA CN ADJ PREP UM CN ADJ PNT REL PREP IND V DA CN REL V PNT
```

Fig. 2 – Exemplo de 1) frase simples; 2) frase anotada; 3) frase composta pelas etiquetas POS

No caso das categorias sintáticas, após uma operação de *parsing* automática[8], usámos as árvores sintáticas e extraímos uma sequência de etiquetas sintáticas através de uma travessia em profundidade.

[8] Após a operação de *parsing* automática os resultados não foram corrigidos manualmente.

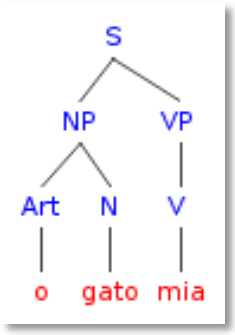

Fig. 3 – *Parsing* da frase “O gato mia” pelo LX-parser.

Na frase exemplificada na Fig.3, obter-se-ia "[S [NP [Art o] [N gato]] [VP [V mia]]]" com a sequência correspondente "S NP Art N VP V".

Para obter os bigramas e trigramas destas unidades sem conteúdo lexical, a ferramenta “CountVectorizer” foi configurada para selecionar bigramas e trigramas, em vez de unigramas, conforme está definido por defeito. Os n-gramas de categorias são modelos de linguagem que permitem obter conjuntos com “n” palavras que ocorram mais frequentemente num determinado *corpus*. No caso dos unigramas, serão os *tokens* mais frequentes, no caso dos bigramas serão os pares de *tokens* mais frequentes, e, no caso dos trigramas, os conjuntos de três *tokens* mais frequentes.

Pontuação e comprimento médio de frase:

Para obter os vetores com as contagens relativas à pontuação, os textos já etiquetados foram processados pelas ferramentas “CountVectorizer” e “TfidfTransformer”. No primeiro caso todos os itens de pontuação foram contabilizados.

Para as contagens de itens lexicais como marcador de autoria foram apenas corridas as ferramentas “CountVectorizer” e “TfidfTransformer” sobre os textos no seu formato não etiquetado.

No caso das contagens de itens lexicais, esta operação permitiu que cada texto fosse transformado num vetor de contagens de cada palavra lexical, com a medida de preponderância aplicada sobre o conjunto dos restantes documentos do grupo.

Para o comprimento médio de frase, utilizou-se a pontuação forte como fronteira de frase, e, após as contagens de cada item lexical, contabilizou-se a média de itens lexicais por frase.

As operações "CountVectorizer" e "TfidfTransformer" foram repetidas para todos os marcadores de autoria, com exceção do cálculo do comprimento médio de frase. Os vetores de cada marcador de autoria foram compilados num vetor único para cada texto, ao qual se adicionou posteriormente a contabilização do comprimento médio de frase.

5.4.1 – Teste I

Atribuição de Autoria

Numa primeira abordagem, pretendemos testar a capacidade da Máquina de Vetores de Suporte de atribuir o autor correto às 12 cartas de ameaça redigidas pelos sujeitos da experiência. Conforme explicámos acima, as Máquinas de Vetores de Suporte permitem apenas atribuir uma classificação binária a partir do estabelecimento de uma margem máxima de separação entre conjuntos definidos de vetores multidimensionais. O facto de o classificador ser binário, enquanto havia 12 autores atribuíveis, levou-nos a optar por um esquema *one-vs-all*, que implica a criação de 12 classificadores binários, um por autor. Dada uma carta, cada classificador está encarregado de decidir se essa carta pertence ao autor associado a esse classificador (uma decisão binária). Caso vários classificadores respondam positivamente, o esquema *one-vs-all* usado permite um desempate baseado no nível de confiança que cada classificador atribui à sua decisão.

Resultados:

O classificador conseguiu atribuir o autor em 58% dos casos. Dos 12 textos questionados, a 7 foi corretamente atribuído o respetivo autor.

Atribuição de Formação Curricular

Para treinar o classificador, os textos foram divididos consoante a instituição onde os sujeitos tinham estudado. O classificador teria assim uma classificação binária entre o *corpus* de treino dos textos dos autores da Faculdade de Ciências e o *corpus* de treino dos textos dos autores da Faculdade de Letras.

Resultados:

O classificador conseguiu atribuir corretamente a Faculdade de origem em 67% dos casos. Dos 12 textos questionados, a 8 foi corretamente atribuída a instituição de Letras ou Ciências.

Atribuição de Género

Para treinar o classificador, os textos foram divididos por género. O classificador teria novamente uma classificação binária, agora entre o *corpus* de textos dos autores do sexo masculino e o *corpus* de textos dos autores do sexo feminino.

Resultados:

O classificador conseguiu atribuir corretamente o género a 92% dos casos. Dos 12 textos questionados, a 11 foi corretamente atribuído o género do seu autor.

5.4.2 – Teste II

Neste segundo teste tentámos reproduzir um caso correspondente ao contexto real judicial. Do conjunto de 12 sujeitos da experiência, escolhemos um a quem pedimos que redigisse uma carta de ameaça adicional, num momento posterior ao da primeira recolha. Este documento foi submetido ao processo de tratamento de texto a que fora submetida a amostra restante, i.e. as 48 cartas do conjunto dos 12 autores. A carta, que inicialmente tinha sido redigida em formato .docx, foi transformada num ficheiro .txt e etiquetada, sintática e morfossintaticamente.

Após executar os processos descritos no capítulo 5.4, correram-se os testes para atribuição de autoria, formação curricular e género, desta vez apenas com uma carta de ameaça – "TextoQ" – enquanto texto questionado. Para atribuição de autoria, cada um dos 12 classificadores, treinado sobre um conjunto de 4 textos de cada autor, classificou o "TextoQ" no esquema "*one-versus-all",* atribuindo autoria ao classificador com o maior valor de confiança. A variável "formação curricular" foi testada com um conjunto de 24 textos de cada uma das duas instituições, usados para treinar o classificador na classificação binária FC/FL. No teste, seria atribuída ao "TextoQ" a pertença ao grupo "FC" (Faculdade de Ciências) ou ao grupo "FL" (Faculdade de Letras). No caso da atribuição de género, o classificador da Máquina de Vetores de Suporte atribuiu um grupo entre os disponíveis na classificação binária M ou F, em que o grupo M foi treinado com o conjunto de textos (24 textos) dos indivíduos do sexo masculino e o grupo F treinado com o conjunto de textos (24 textos) dos indivíduos do sexo feminino.

Estes foram os resultados obtidos na realização do Teste II::

Variável	"TextoQ"	Variável Atribuída	Taxa de acerto
Autoria	FA	FA	100%
Formação Curricular	FC	FC	100%
Género	M	M	100%

Tabela 2 – Resultados para atribuição de autoria, formação curricular e género ao "TextoQ"

5.4.3 – Discussão dos resultados

Os resultados obtidos para o Teste I permitiram aferir a capacidade de a SVM classificar adequadamente um texto de acordo com as possibilidades disponíveis para género, formação curricular e autor. Em cada uma das classificações, a taxa de sucesso deve ser adequadamente analisada, de acordo com o número de opções disponíveis. Por exemplo, uma taxa de sucesso de 57% para atribuição de autoria será mais significativa do que uma taxa de sucesso de 68% para atribuição de formação curricular, considerando que para a primeira operação havia 12 opções disponíveis (1/12 = 0,083(3) de probabilidade de sair aleatoriamente o autor correto), comparativamente às duas opções disponíveis na segunda operação (1/2 = 0,50 de probabilidade de sair aleatoriamente a Faculdade correta). Nesse sentido, julgamos mais relevantes os resultados obtidos para a atribuição de género e de autor, e menos relevantes os conseguidos para a atribuição de formação curricular. Para o caso do género, aliás, há já vários estudos (Mouton (2000), Chesire (2002), Pérez (2007)) que destacam as diferenças no comportamento linguístico de homens e mulheres.

Na concretização destes testes é também possível obter os valores de confiança para cada uma das classificações (ANEXO I). Estes valores de confiança são, na realidade, a distância do novo ponto classificado em relação ao hiperplano. O classificador opta pelo texto que apresenta a maior distância em relação ao hiperplano, o que indicará uma maior proximidade em relação ao *corpus* do autor correto. No entanto, considerando a taxa de acerto geral do classificador (cf. 5.4.1), pode ser calculada a taxa de acerto, não para um autor específico, mas para um conjunto de autores que o classificador considera serem os mais prováveis. Este cálculo intitula-se "top-N accuracy" em que a "N" corresponde o número de autores a que o classificador atribui a autoria, com as respetivas taxas de acerto. O classificador permite assim identificar, não apenas o mais provável autor de cada carta (top-1 accuracy), mas também o conjunto dos autores mais prováveis de cada carta.

Estes são os resultados do classificador para os seguintes conjuntos de autores considerados:

Top-N accuracy	Cartas corretamente atribuídas	Taxa de acerto
Top-1 accuracy	7/12	58%
Top-2 accuracy	8/12	67%
Top-3 accuracy	11/12	92%
Top-4 accuracy	12/12	100%

Tabela 3 – Top N-accuracy para o classificador da SVM

A utilização desta medida de confiança permite-nos reduzir o conjunto de autores possíveis. Com a redução de um conjunto de 12 para 4 autores, a análise qualitativa pode ser mais rigorosa e permitir uma apreciação linguística mais cabal, aproximando-se ao mesmo tempo dos contextos reais de peritagem linguística, em que existe normalmente um conjunto muito limitado de autores suspeitos. Conforme Coulthard (2004:2): "Thus, the task of the linguistic detective is never one of identifying an author from millions of candidates on the basis of the linguistic evidence alone, but rather of selecting (or of course *deselecting)* one author from a very small number of candidates, usually fewer than a dozen and in many cases only two (Coulthard 1992, 1993, 1994a, b, 1995, 1997, Eagleson, 1994)."

Embora os métodos computacionais confiram uma maior fiabilidade ao processo de atribuição de autoria, ainda é precoce assumir exclusivamente uma abordagem quantitativa para a peritagem linguística. Conforme afirmado anteriormente, uma combinação de métodos quantitativos e qualitativos continua a ser a metodologia preferencial (Marquilhas e Cardoso (2011:418), Litosseliti (2010:50)). Nesse sentido, os resultados conseguidos nestes testes com a Máquina de Vetores de Suporte são uma contribuição positiva e indiciam que este possa ser um bom método de delimitação de autores, embora seja recomendável testá-lo mais exaustivamente com outro tipo de *corpora*, outro tipo de marcadores de autoria e amostras mais variadas.

5.5 – Análise Qualitativa

Os elementos linguísticos marcados de um estilo idioletal, em teoria, corresponderão às escolhas do autor para aquela produção textual, sem esquecer as variações intrínsecas ao registo e ao género textual. Conforme afirmado por Almeida (2014:157): "nas abordagens estilísticas, não se propõe que apenas um ou outro marcador seja utilizado sempre, independentemente do caso, como um universal, mas sim que cada indivíduo apresente um conjunto de características que o identifique, e este conjunto pode variar entre indivíduos."

Na história da linguística forense, há alguns casos conhecidos que tiveram a ver com o isolamento de marcadores estilísticos, fruto de análises qualitativas. Conforme Coulthard (2006:2) explica, foi esse o caso do "Unabomber", que em 1995 foi identificado devido a uma expressão multipalavra reconhecida pelo seu irmão como típica da sua "terminologia", ou vocabulário idiossincrático. O FBI contrastou o manifesto de 35.000 palavras de Ted Kaczynski com um artigo do mesmo autor de 300 palavras escrito uma década antes, e atestou bastantes similaridades, listando um conjunto específico de palavras lexicais, gramaticais e algumas expressões fixas. A defesa contratou uma linguista que argumentou que qualquer pessoa poderia usar o conjunto de itens destacado e que o vocabulário partilhado não poderia ter assim tanto significado. Porém, uma pesquisa na internet do conjunto específico de itens elencados reunidos num só documento apenas devolveu 65 resultados, todos estes pertencentes a versões do manifesto do mesmo autor, Ted Kaczynski (Coulthard (2006:3)).

Em Turell (2010:227), num caso de atribuição de autoria a um conjunto de emails com mensagens de extorsão, foi feita uma recolha de itens estilísticos para verificar a possibilidade de aproximar o estilo idioletal do autor dos textos questionados ao estilo do texto de um conjunto de faxes de autoria conhecida. Neste caso, os fenómenos linguísticos que foram isolados, e que eram fenómenos típicos de línguas em contato, nomeadamente o Catalão e o Espanhol, contribuíram para a identificação do autor dos textos de extorsão questionados.

Nesta dissertação, ao analisarmos o "TextoQ", tentámos elaborar um perfil linguístico baseado nos elementos que se destacam como provavelmente marcados. Em relação ao *corpus* que compilámos para a realização dos testes experimentais, considerámos agora só um conjunto mais reduzido de autores, apenas quatro, como autores possíveis para o "TextoQ" – a carta de ameaça questionada, de acordo com os resultados obtidos a partir dos valores de confiança (ANEXO I). Assim foi possível circunscrever o número de documentos a 16 (em vez dos 48 iniciais), o que permitiu uma análise comparativa mais praticável. Incidiu sobre o seguinte conjunto de quatro autores: DM, FA, JC e JV

Consideremos o "TextoQ":

É esperado algum decoro de uma figura pública. Fazes as tuas aparências, dás a cara por umas coisas bonitas, mostras as namoradas... Tudo bem. Ninguém disse que tinhas de ser um segundo Gandhi, defender as criancinhas todas e acabar com a fome no mundo. Mas há limites. E cruzaste o limite quando começaste a falar mal de pessoas que não conheces, quando começas a ser injusto sobre o que não conheces, e quando começas a fazer mal sem justa causa. E foi mesmo isso que fizeste na teu último discurso sobre a nossa cidade.

A tua vida é pública – tu fazes questão de a tornar pública. Por isso sabe-se perfeitamente que nunca moraste aqui, mal passaste por cá, e as tuas vivências com as pessoas desta terra são muito limitadas. Por isso cada vez menos se percebe qual a tua embirrança com a nossa cidade. Este é um local pacífico, pacato, de boas gentes e bons costumes. Sim, não somos ricos, nem perto disso. Mas a riqueza não mede o coração das pessoas. E por isso as tuas palavras sobre sermos um esgoto cheio de bandidos corruptos e uma amolgadela civilizacional são das coisas mais injustas ditas em público. E se calhar por isso, vais ter razão.

Espero que peças desculpa pelo teu último discurso da próxima vez que falares em público, ou as tuas palavras sobre o carácter dos habitantes da nossa cidade poderá ter uma pintinha de razão. O teu carro poderá aparecer partido, ou a casa roubada, ou até quem sabe não acabas com uns ossos partidos para teres uns tempos de meditação no hospital. Lá poderás ter alguma paz para pensar sobre a natureza humana. Ou não. Os hospitais não costumam ter muita segurança à noite. Agora tu é que sabes, se preferes ficar com a tua difamação e umas quantas contas de hospital, ou corrigir os teus erros e viveres em paz.

Fig. 4 – Carta de ameaça considerada para "TextoQ"

Para o levantamento de marcadores estilísticos, procedeu-se a uma leitura cuidada do texto questionado e isolaram-se aqueles elementos lexicais, sintáticos e ortográficos que nos pareceram mais individualizantes, bem como certos recursos expressivos associados à retórica, i.e., figuras de retórica ou tropos. Neste processo, pretendemos identificar o estilo idioletal do autor do "TextoQ" e detetar esses mesmos traços em algum dos quatro autores disponíveis.

Concordamos que é desafiante conseguir estabelecer o contraste entre o que é marcado, ou saliente, e o que é considerado neutro numa determinada língua. Em princípio, o mais neutro será o padrão, dada a sua menor variabilidade. Na gramática de Cunha e Cintra (1984), o conceito de língua-padrão surge definido como "uma entre as muitas variedades de um idioma, [mas] é sempre a mais **prestigiosa**, porque actua como modelo, como norma, como **ideal linguístico** de uma comunidade (…)". Nesta perspetiva, as escolhas que podemos considerar não marcadas no estilo idioletal de um falante seriam as que mais se aproximam do que aparece prescrito em gramáticas, dicionários e prontuários. No entanto, é preciso também ver que nem sempre esta norma assim definida, precisamente pelo prestígio de que é investida, será a que ocorre mais frequentemente. A verdade é que é preciso distinguir entre dois modelos de língua, implícitos na definição de Cunha e Cintra mas explícitos em muitos trabalhos de sociolinguística. Trata-se do modelo da norma culta ("prestigiosa") e do modelo da norma padrão (um "ideal linguístico"). O primeiro é real, mas exclusivo; o segundo é mais geral, mas imaginário, implicando uma unicidade que nunca se poderá verificar no uso natural das línguas (Mateus e Cardeira 2007: 22).

Como o que se torna relevante no âmbito da linguística forense é um instrumento que permita ao investigador isolar os aspetos marcados que configurem um estilo idioletal, é preferível lidar diretamente com o *uso* da língua e a sua inerente variação, até porque dispomos, hoje em dia, de *corpora* textuais de grande dimensão. Com eles tornou-se possível formar uma ideia, para cada comunidade linguística, sobre quais são os comportamentos típicos, logo, não marcados, dos falantes e escreventes da língua em causa. Porque permitem pesar a representatividade de determinadas opções sintáticas, lexicais,

retóricas e ortográficas, também permitem isolar as formas mais marcadas no uso da língua, que serão, simultaneamente, as de frequência mais rara neste tipo de recursos.

Parecer linguístico

No universo de quatro autores considerados suspeitos de terem escrito a carta de ameaça, pretendemos encontrar uma resposta para esta questão: "Qual dos quatro autores considerados escreveu o "TextoQ"?

5.5.1 – Análise qualitativa do texto questionado

O texto questionado apresenta um conjunto de marcadores que iremos tratar de forma sequencial. Após a sua listagem e descrição, tentaremos compreender em que medida cada um destes itens é marcado, e, posteriormente, como se manifesta no conjunto de textos de que dispomos para cada um dos autores suspeitos.

Para verificarmos se uma determinada palavra ou estrutura sintática era frequente, ou normal, no uso da língua, usámos o Corpus de Referência do Português Contemporâneo (CRPC)[9], por ser um *corpus* de grande dimensão (c. de 3 milhões de palavras para a variedade Português de Portugal, que foi a utilizada nesta análise) e bastante diversificado (inclui textos literários, jornalísticos, técnicos, didáticos, jurídicos, etc.).

Segue-se o levantamento das estruturas que julgámos serem marcadas para o texto considerado:

(1) *Passiva sintática impessoal em início de frase.* No início da carta de ameaça considerada como "TextoQ" o autor utiliza, para obter um efeito de indeterminação do sujeito[10], uma construção passiva sintática impessoal, "É esperado algum decoro" [linha 1], ao invés de optar pela forma passiva de *-se* impessoal, "Espera-se algum decoro", que é para o

[9] Corpus de Referência do Português Contemporâneo disponível em www.clul.ul.pt
[10] Cf. Cunha e Cintra (1984:150).

português, bem como para as restantes línguas românicas de sujeito nulo, uma construção frequente (Duarte (2003: 532). A pesquisa no CRPC das duas estruturas em posição inicial de frase, mantendo os verbos no mesmo tempo, modo e pessoa que os considerados, provou haver uma clara prevalência da passiva de *-se* impessoal (1162 ocorrências) sobre a passiva sintática impessoal (8 ocorrências).

(2) *Impropriedade vocabular/ erro de seleção semântica.* A construção frásica em consideração, "Fazes as tuas aparências" [linha 1], é irregular em português, uma vez que o nome "aparências" tem restrições de seleção semântica que não podem coocorrer com o verbo "fazer", embora uma pesquisa no CRPC devolva resultados para construções como "fabricar (aparências)" ou "criar (aparências)". Por outro lado, a estrutura "Fazes os teus aparecimentos", i.e "aparecimentos (em público)", também seria possível, o que nos leva a assumir que o autor poderá ter cometido um erro de impropriedade vocabular, substituindo "aparecimentos" por "aparências".

(3) *Erros de concordância*. O texto manifesta alguns erros de concordância verbal e nominal. Existe falta de concordância entre a oração principal, "cruzaste o limite" [linha 4], com o verbo no pretérito perfeito, e as duas últimas orações da estrutura frásica que empregam o verbo no presente do indicativo "quando começas" [linhas 5 e 6]. Também na construção "na teu último discurso" [linha 7] observamos um erro de concordância nominal, uma vez que o determinante artigo definido feminino da contração da preposição "em + a" não concorda em género com o nome masculino que atua como núcleo do sintagma nominal a que pertence: "discurso". Também verificamos falta de concordância sujeito-verbo na sequência "as tuas palavras sobre o carácter da nossa cidade poderá" [linha 19], uma vez que o sujeito é plural mas o verbo se apresenta no singular.

(4a) *Figuras de retórica ou tropos: amplificação por anadiplose, anáfora e epístrofe.* Podemos observar uso de amplificação por anadiplose quando o autor recorre à palavra "limites" [linha 4], no final da frase "Mas há limites", retomando a mesma palavra no início da frase consecutiva, "E cruzaste o limite quando (...)". Vemos também amplificação por anáfora nas sequências iniciadas por "quando" [linhas 4-6]: "quando começaste a falar mal", "quando começas a ser injusto" e "quando começas a fazer mal", bem como nas frases

iniciadas por "e": “E cruzaste o limite quando começas a falar mal” [linha 4], “e quando fazes o mal sem justa causa” [linha 6], e “E foi mesmo isso que fizeste” [linha 6]. Também observamos esta estratégia estilística no uso que o autor faz do marcador discursivo “por isso” [linhas 9, 11, 14 e 16], que utiliza recorrentemente no texto. O autor recorre ainda à amplificação por epístrofe quando reutiliza a mesma palavra, “pública”, para finalizar as duas orações consecutivas “a tua vida é pública” e “tu fazes questão de a tornar pública” [linha 9];

(4b) *Figuras de retórica ou tropos: ironia.* O autor serve-se regularmente da ironia. Esta figura de linguagem manifesta-se por permitir obter, a partir do contexto do enunciado, um “significado literal que diverge ou é mesmo contraposto ao significado que corresponde à intenção do emissor e que o receptor pode e deve interpretar mediante a análise do co-texto e sobretudo do contexto”, conforme lemos no Dicionário Terminológico (2015). O autor do “TextoQ” ameaça o destinatário de forma indireta e disfarça esta intenção sob a forma de elogio: “as tuas palavras sobre o carácter dos habitantes da nossa cidade [poderão] ter uma pintinha de razão”, servindo-se do verbo “poder” com modalidade epistémica ou modalidade externa (Oliveira e Mendes (2013: 644)). No entanto, é compreensível pelo contexto que é na realidade uma modalidade de ironia, por se tratar de uma intimação para que o interlocutor cumpra as exigências do autor, ameaçando-se sob a capa de um elogio. Também observamos ironia na forma como a ameaça aparece disfarçada de promessa positiva: “quem sabe não acabas com uns ossos partidos para teres uns tempos de meditação no hospital”.

(5) *Pontuação.* No texto questionado ocorrem os sinais de pontuação mais comuns, vírgula e ponto final, mas também o travessão, que se manifesta menos frequentemente: “A tua vida é pública – tu fazes questão de a tornar pública.” [linha 9]. De acordo com o Dicionário Terminológico (2015), o travessão é usado como sinal de pontuação para intercalações de palavras ou frases. No entanto, nesta carta de ameaça o travessão introduz uma frase que não tem valor parentético, mas sim de conclusão (Cunha e Cintra (1984:663)), atuando como alternativa ao uso de dois pontos. Por representar uma opção ortográfica do autor, consideramos a sua utilização como um possível marcador de autoria textual.

(6) *Organização textual.* A um nível suprassintático, há ainda que considerar que a organização das sequências textuais pode conter traços idiossincráticos que contribuam para a caracterização do estilo do autor. Com efeito, o estilo é 'um conjunto global de traços recorrentes do plano do conteúdo (formas discursivas) e do plano da expressão (formas textuais), que produzem um efeito de sentido de identidade'. (Fiorin (2008:97) apud Almeida (2014:164))".

O texto questionado tem sequências textuais de vários tipos, mas prevalecem as sequências argumentativas (Adam (1992)). Neste tipo textual, uma tese (ou argumento) é fundamentada em premissas para conduzir o interlocutor à aceitação de uma conclusão. O autor do texto Q estrutura a sua argumentação com o avanço de uma premissa, "É esperado algum decoro de uma figura pública" [linha 1], e seguidamente apresenta, de forma sequencial, as razões pelas quais entende que o interlocutor não está a responder à expectativa anunciada. Justifica desta forma a ameaça que faz ao interlocutor: "Espero que peças desculpa pelo teu último discurso da próxima vez que falares em público, ou as tuas palavras [poderão] ter uma pintinha de razão." [linha 18], que é uma ameaça sob forma de elogio (cf. com ponto (4b)).

5.5.2 – Texto questionado *vs.* textos da amostra

Após o levantamento dos traços acima identificados, virámo-nos para o conjunto das cartas dos quatro autores indicados como prováveis pelo teste "*top-N accuracy*" da Máquina de Vetores de Suportes, i.e., DM, FA, JC e JV. Dispúnhamos de quatro cartas para cada autor, com cerca de 300 palavras para cada uma: uma carta de ameaça, uma carta de extorsão, uma carta de agradecimento e uma carta de reclamação.

(1) *Passiva sintática impessoal em início de frase.* No conjunto das cartas questionadas, não observamos nenhuma construção de passiva sintática impessoal em início de frase.

(2) *Impropriedade vocabular/ erro de seleção semântica.* A única ocorrência relevante encontra-se na carta de extorsão do informante FA: "sempre confirmou que o senhor é fiel, e que nunca haveria terceiras rodas no vosso casamento". Não conseguimos encontrar ocorrências para a expressão "terceiras rodas" que fossem semanticamente adequadas ao contexto. No entanto, encontramos bastantes ocorrências no CRPC para as expressões "segundas rondas" e "terceiras rondas". Dado que "terceiras rondas" seria sinónimo de "terceiras voltas", a expressão tem de ser analisada como um caso de impropriedade vocabular, com a palavra "roda" a ser usada em lugar de "ronda".

(3) *Erros de concordância*. Os textos manifestam algumas faltas de concordância. Encontramos uma falta de concordância verbal na carta de agradecimento do informante JV: "Somos muito próxima". O grupo adjetival que constitui o predicativo do sujeito está no singular, quando o verbo copulativo surge na primeira pessoa do plural. Na carta de ameaça do informante FA, encontramos também uma falta de concordância sujeito-verbo: "a minha meditação e paz na paróquia seja interrompida". Este sujeito composto "a minha meditação e paz na paróquia" requer que o verbo seja conjugado na terceira pessoa do plural, i.e., "sejam", e não "seja".

(4a) *Figuras de retórica ou tropos: amplificação por anadiplose, anáfora e epístrofe.* Não identificamos no conjunto das cartas dos quatro informantes recurso à amplificação por

anadiplose. No entanto, os fenómenos de amplificação por anáfora estão presentes nos quatro informantes considerados: DM, FA, JC e JV.

O informante DM apresenta amplificação por anáfora na carta de ameaça, com repetição do verbo *ir*: "Vais pagar por aquilo que me tens feito e vais ter de começar do início(...)", "Vais provar do teu próprio veneno". Repete também o pronome relativo "que": "Nós, que nunca te quisemos mal, que te ajudámos, que te apoiámos sempre". Na carta de agradecimento, este informante também usa seguidamente "Agradeço": "Agradeço a tua paciência (...)" e "Agradeço por me teres apoiado sempre (...)".

O informante FA, na sua carta de agradecimento, serve-se frequentemente desta figura de linguagem: "Que outros mundos (...)", "Que planetas (...)" e "que outras criaturas"; também na repetição da conjunção *e*: "e eu acompanho-o, e alimento-o. (...)"; num outro momento do texto, a amplificação com recurso ao pronome relativo "que": "que me acompanharam", "que me ensinaram". Na carta de agradecimento do informante JC observamos também uma amplificação por anáfora com a palavra *nunca*: "Se outros ajudou (...), nunca isso se notou, nunca a sua atenção a outra pessoa significou uma desatenção para comigo", e com o advérbio *só:* "Só assim a minha gratidão será consumada, só assim estas palavras serão verdadeiramente uma carta que alcança o seu escopo".

O informante JV também recorre à amplificação por anáfora na carta de agradecimento, iniciando vários períodos enunciativos, sequenciais e não sequenciais, com a estrutura "agradeço": "Agradeço todo o carinho", "Agradeço por todas as vezes em que me ouviram", "Agradeço por me apoiarem", "Agradeço a dedicação", "Agradeço também aos meus avós", etc.. Este informante também se serve de amplificação por anáfora na sua carta de extorsão: "Consegue imaginar a sua vida sem o luxo (...)", "Consegue imaginar a sua vida sem a sua carreira(...)", "Consegue imaginar-se preso?" e "Consegue imaginar-se sem amigos (...)?"

Quanto à amplificação por epístrofe, o informante FA recorre a essa modalidade de amplificação ao usar a palavra *obrigado* como elemento finalizador dos três parágrafos finais da sua carta de agradecimento.

(4b) *Figuras de retórica ou tropos: ironia.* Os informantes que recorrem à ironia são os informantes JC e FA. O informante JC apresenta ironia na sua carta de extorsão quando pede uma "soma simpática de dinheiro" em troca da não divulgação de informações críticas, informando que irá prejudicar a carreira do destinatário, caso este, segundo as suas

palavras, “não tenha a gentileza de me julgar um seu comparsa, e mesmo quase amigo”. O informante FA também recorre a esta figura de linguagem em três das suas cartas: a de ameaça, a de extorsão e a de reclamação. Na carta de ameaça, fá-lo referindo-se ao seu interlocutor que ‘agraciando os restantes paroquianos com a sua presença’ deixa a caixa das doações “ligeiramente mais pobre” após a sua passagem por ela. Na carta de extorsão, este informante volta a recorrer a este recurso estilístico quando defende que “há oportunidades no mundo e que devemos todos beneficiar com elas” e que neste caso foi “oportuno ter visto e fotografado” o interlocutor com “uma senhora que não é a sua esposa” a entrar num “motel onde os quartos são alugados à hora”. Reforça o tom irónico quando menciona que estará disposto a ignorar “qualquer conhecimento ou prova que possua acerca do tão feliz encontro descrito acima”, uma vez que não deseja “arruinar a felicidade de ninguém”. Termina por reforçar este registo de ironia dizendo: “Será um prazer fazer negócios consigo”. Voltamos a observar esta figura de linguagem na sua carta de reclamação, ao se referir a um produto que lhe foi entregue danificado como sendo “um pisa-papéis caro”.

(5) *Pontuação.* Apenas dois dos quatro informantes da amostra escolheram usar travessão nos seus textos, os informantes JC e FA.

O informante JC usa este sinal de pontuação nas três possibilidades descritas para o seu uso: intercalação de palavra ou frase, “se outros ajudou – e bem sabemos que sim! –, nunca isso se notou”; início de um enunciado em discurso direto, “– Exagero! – responderia sem demora”; e, conforme se manifesta no texto questionado, como introdução de uma conclusão, “caiam nas mãos erradas – da polícia, por exemplo, ou daquela empresa que há tanto tempo o senhor tenta vencer”.

Também no caso do informante FA o travessão é usado para introduzir este valor de conclusão, conforme observamos na carta de ameaça: “para que todos saibam o que andas a fazer à comunidade – as tuas acções não podem continuar impunes!”. Encontramos também por duas vezes o travessão como sinal de pontuação sinalizando uma palavra ou frase intercalada, uma vez na carta de ameaça: “para que possamos continuar a ajudar os mais necessitados são necessários bens materiais – dinheiro – já que nada neste mundo é grátis” e uma segunda vez na carta de extorsão “uma senhora que – perdoe o fácil julgamento – deve alugar o seu “amor” à hora”.

(6) *Organização textual.* Considerámos o "TextoQ" predominantemente argumentativo. Os textos do conjunto dos quatro autores com caráter mais argumentativo, isto é, que expõem uma premissa inicial seguida de argumentos para levar o interlocutor a aceitar uma conclusão, são as cartas de reclamação. No caso das cartas de ameaça, verificámos que apenas o informante FA adota o tipo predominantemente argumentativo, orientando raramente o seu discurso ao interlocutor (tipo de texto dialogal), fazendo-o apenas já no final da carta e para formalizar a ameaça. Os restantes informantes, DM, JC e JV, todos escolheram um tipo de texto essencialmente dialogal na composição das suas ameaças. No caso das cartas de extorsão, o informante DM é essencialmente dialogal, os informantes JC e JV optam pelos tipos textuais narrativo e dialogal, enquanto FA apresenta um texto distinto, com sequências argumentativas, narrativas e dialogais alternadas. Quanto às cartas de agradecimento, o informante DM opta por um texto essencialmente dialogal. O informante JC apresenta um texto bastante rico em termos tipológicos, inicialmente narrativo, por vezes descritivo e dialogal. Este informante demarca-se dos restantes não apenas pela diversidade de sequências textuais, mas também pela riqueza de vocabulário. O informante JV opta por um texto pleno em sequências declarativas aproximando-o mais do tipo de texto expositivo-explicativo. Já o informante FA apresenta um texto com sequências narrativas e argumentativas, finalizando com sequências textuais declarativas, o que o aproxima mais do texto expositivo-explicativo.

Conclusões:

Após a análise comparativa do "TextoQ" com os textos dos informantes DM, FA, JC e JV, concluímos que o informante FA apresenta uma concentração superior dos marcadores autorais do texto questionado na sua amostra textual.

5.6 – Discussão das conclusões da análise combinada

Em atribuição de autoria, a análise qualitativa da peritagem linguística depende da coocorrência de um conjunto de marcadores de autoria, conforme Marquilhas e Cardoso (2011:427): "(...)na análise qualitativa desenvolvida pela linguística forense, a singularidade enunciativa não é estabelecida habitualmente pela presença de um marcador de estilo, mas sim pela coexistência de vários marcadores nos mesmos grupos de texto." A análise que foi levada a cabo neste estudo contemplou um conjunto de marcadores de autoria que foram devidamente circunscritos e que considerámos idiossincráticos do autor do "TextoQ". Tal como argumentámos no Capítulo 3, o conjunto de escolhas idiossincráticas de um autor pode contribuir para definir não a sua "impressão digital", mas o seu estilo idioletal, o que ajuda a identificar um autor certo ou, pelo menos, a eliminar autores que não correspondam ao perfil encontrado.

Apesar de os textos da amostra pertencerem ao mesmo género textual, conseguimos identificar diferenças linguísticas que isolam os textos dos informantes DM, JC e JV do texto questionado. Uma análise mais exaustiva poderia passar por elaborar um perfil linguístico não apenas do autor do texto questionado, mas também de cada autor das sucessivas amostras consideradas. Porém, à semelhança dos contextos forenses reais, o nosso foco incidiu sobre um texto questionado e as suas caraterísticas distintivas, não sobre a produção textual de um universo de suspeitos[11].

Os resultados obtidos nesta análise parecem confirmar os que se obtiveram com o classificador da Máquina de Vetores de Suporte. No conjunto dos testes realizados, verificámos ser possível atribuir a autoria correta ao texto questionado com alguma margem de confiança[12], uma vez que as outras amostras pertenciam ao mesmo género textual e, ainda assim, o estilo idioletal do autor do "TextoQ" manifestou sempre algum contraste.

[11] Cf. Owen Amos, "The Text Trap" em *The Northern Echo* (visitado em 20 de Novembro de 2015, http://www.thenorthernecho.co.uk/news/2076811.print/)

[12] Cf. N. de rodapé 13.

6 – Notas conclusivas

Com esta dissertação pretendemos salientar as vantagens de uma análise combinada para atribuição de autoria em linguística forense. Para cumprir esse objetivo, elaborámos uma experiência que pretendia, em primeiro lugar, testar o método quantitativo sobre o *corpus* reunido, usando amostras textuais dos próprios autores para verificar a probabilidade de acerto da máquina, e, em segundo lugar, simular um caso de atribuição de autoria verosímil para o contexto forense. Na constituição do *corpus* tentámos controlar variáveis tais como o género e a formação curricular, eliminando ao mesmo tempo o efeito de outros fatores de variação linguística, tais como a atividade profissional, o dialeto e a faixa etária (cf. 5.2).

Na escolha de marcadores de autoria na análise quantitativa, selecionámos os métodos que pareciam apresentar resultados mais significativos na bibliografia de linguística forense que dá conta de experiências anteriores bem sucedidas. Daí a importância dada por nós à dimensão do texto questionado e ao *corpus* para treino do classificador.

Os resultados a que chegámos com o classificador da Máquina de Vetores de Suporte indicam que é possível isolar um conjunto de autores possíveis de entre um universo mais amplo de sujeitos, de forma a se poder prosseguir com maior segurança para uma segunda modalidade de análise, já qualitativa. Esta centrou-se no Texto Q e no seu contraste marcado com o uso da língua portuguesa tal como é intuído pelo linguista e confirmado por medições num corpus de referência de grandes dimensões.

Da experiência global, i.e., dos resultados das análises quantitativa e qualitativa, surgiu sempre o mesmo sujeito, o informante FA, como o possível autor do texto questionado, embora a taxa de acerto inicial do classificador ficasse em 58%[13]. Com efeito, no Teste I, o classificador só conseguiu atribuir a 7 das 12 cartas de ameaça o seu autor respetivo. Um dos casos do sucesso na atribuição de autoria verificou-se ser o autor FA, que foi o autor do "TextoQ" no Teste II.

[13] Para um *baseline* de 8,3%. Acerto de 100% em *top-4-accuracy.*

Em análises futuras, pretenderemos alargar o conjunto dos marcadores de autoria de forma a que, gradualmente, possamos verificar o impacto de cada um deles na taxa de acerto do classificador, considerando que uma maior quantidade de traços estilísticos reunidos parece contribuir para uma maior taxa de sucesso (Hirst e Feiguina (2007), Sousa-Silva *et al.* (2005), De Vel *et al.* (2001)).

A taxa de acerto do classificador quanto à formação curricular pareceu pouco significativa, o que poderá denotar a não existência de elementos linguísticos suficientemente distintivos para esta atribuição, pelo menos no contexto desta experiência. Acreditamos, por isso, que as variações estilísticas poderão ser ditadas por fatores sociais mais preponderantes do que o da mera formação curricular.

Já os resultados conseguidos para atribuição de género são indicativos da possibilidade de isolar elementos que permitam a distinção entre a escrita de homens e de mulheres, como aliás vem sendo indicado pelos estudos de sociolinguística, e também de linguística computacional, desenvolvidos nos últimos anos que incidem sobre tais diferenças[14]. Cremos ser possível vir a demarcar de forma mais descritiva em que medida se diferencia o discurso dos homens do discurso das mulheres, também para o caso do Português Europeu, até porque os testes quantitativos ajudam a elaborar melhor a base de tal diferenciação.

A análise qualitativa do texto questionado permitiu-nos isolar traços estilísticos diferenciadores em relação aos textos suspeitos do conjunto dos informantes DM, FA, JC e JV, considerados como conjunto mínimo para uma taxa de acerto de 100% no teste "top-N accuracy". O conjunto dos traços isolados permitiu fazer o levantamento das ocorrências dos mesmos fenómenos linguísticos em poucos textos dos quatro autores. O autor com uma manifestação mais consolidada do conjunto dos traços reunidos foi o informante FA, que corresponde ao autor correto, embora se possa admitir que o "TextoQ" tinha uma dimensão invulgar, já que muitas cartas de ameaça são compostas por apenas algumas frases. Ainda assim, a reunião destas características linguísticas confirma a teoria de que as escolhas que

[14] Cf., por exemplo, Mouton (2000), Chesire (2002), Koppel *et al.* (2002) e Pérez (2007).

um falante faz de forma consistente no quadro do seu sistema linguístico contribuem para delinear o seu estilo idioletal, estilo esse que se manifestará nos enunciados que produz, os quais denunciam, assim, a identidade do autor.

Anexo I – Tabela com os valores de confiança para Teste I e Teste II

Valores mais altos implicam maior confiança.

Atribuição de autor ao TextoQ - Confiança do classificador para cada um dos 12 autores.

Teste II

AC	AN	BH	DM	DO	FA	JC	JV	MG	NB	PO	SA	*gold*	Certo?
-1,00839	-0,86479	-0,92883	-0,80178	-0,95457	-0,73887	-0,77081	-0,74354	-0,83525	-0,85977	-0,85671	-0,81334	*FA*	**sim**

Atribuição de autor às cartas de ameaça - Confiança do classificador para cada carta (linhas) e para cada um dos 12 autores (colunas).

Teste I

AC	AN	BH	DM	DO	FA	JC	JV	MG	NB	PO	SA	*gold*	certo?
-0,56256	-0,80210	-0,85037	-0,80181	-0,88042	-0,93918	-0,79676	-0,85029	-0,84104	-0,75045	-0,95142	-0,94020	*AC*	**sim**
-0,85850	-0,74596	-0,99828	-0,76759	-0,80313	-0,67425	-0,87400	-0,85551	-0,88829	-1,06135	-0,67034	-0,82907	*AN*	não
-0,63440	-0,89566	-0,73943	-0,79381	-1,06794	-0,94703	-0,62777	-0,99483	-0,84114	-0,81533	-0,69638	-0,84225	*BH*	não
-0,78715	-0,80255	-1,10975	-0,60769	-0,68419	-0,86608	-0,79132	-0,80714	-0,69985	-0,83477	-0,87454	-1,01992	*DM*	**sim**
-0,89066	-0,90602	-0,88770	-0,68150	-0,45133	-0,72834	-0,81332	-0,86221	-0,97468	-0,96406	-0,76869	-0,94042	*DO*	**sim**
-0,95635	-0,78188	-0,99123	-0,73324	-0,63932	-0,63221	-0,90239	-0,86976	-0,80155	-0,85944	-0,99627	-0,84028	*FA*	**sim**
-0,86564	-0,67250	-0,93757	-0,69803	-0,67985	-0,88861	-0,66673	-1,06864	-0,84104	-0,81773	-0,75443	-0,86087	*JC*	**sim**
-0,94680	-0,87315	-0,94449	-0,74131	-0,65444	-0,81907	-0,94794	-0,51720	-0,80867	-0,90122	-0,73307	-0,89336	*JV*	**sim**
-0,69333	-0,74922	-0,83209	-0,77948	-0,76536	-0,78154	-0,77030	-0,82107	-0,73009	-0,92337	-0,62228	-0,80625	*MG*	não
-0,87891	-0,96633	-0,86026	-0,87080	-0,74607	-0,94918	-0,73559	-0,70536	-0,85577	-0,73691	-0,77649	-0,84160	*NB*	não
-0,78744	-0,67057	-1,00610	-0,77834	-0,38067	-0,61540	-0,82258	-1,01127	-0,69824	-0,92760	-0,41563	-1,03169	*PO*	não
-0,93356	-0,95210	-0,80113	-0,78689	-0,85736	-1,09099	-0,74508	-0,68084	-0,94939	-0,70940	-0,89578	-0,37503	*SA*	**sim**

Anexo II – Amostras textuais dos quatro autores suspeitos

Informante DM

Agradecimento

Sei que não esperavas ler esta carta, mas não sou muito boa com as palavras faladas e, por isso, tento fazê-lo com as palavras escritas. Tenho que agradecer tudo o que tens feito por mim. Agradeço a tua paciência, por vezes infinita, para me aturar nos momentos mais complicados da minha vida. Agradeço por me teres apoiado sempre, mesmo quando a pessoa que precisava de apoio eras tu. Obrigada por me teres feito sentir a pessoa mais inteligente, mais capaz, bonita e engraçada, até nas alturas em que me senti tudo menos isso. Obrigada por me teres segurado na mão tantas vezes, por teres estado ao meu lado e ajudado a levar o meu barco a bom porto. Sei que estás comigo porque queres, porque escolheste e porque fizeste um compromisso comigo. Obrigada por me escolheres e por teres partilhado comigo a tua vida. A vida não tem sido fácil para nenhum dos dois, mas torna-se mais leve quando há mais alguém para suportar o fardo. Quando há alguém para nos apoiar, ajudar, para nos fazer rir e para oferecer um ombro onde chorar. Tu tens feito sempre isso sem pedir nada em troca. Penso que não sabes o quão és importante para mim e o quanto me ajudaste a crescer e a evoluir. Só espero que continues a fazer esta caminhada comigo, porque temos ainda tanto caminho para trilhar. Agradeço todos os dias por ter alguém como tu a meu lado e espero que nunca me faltes, tal como eu espero nunca faltar contigo e conto estar sempre a teu lado. Obrigada por seres a pessoa calma nos momentos mais ansiosos e evitares que entre em pânico. Por seres a voz da razão e por me conseguires dar outra perspectiva face aos meus problemas. Por relativizares e por me fazeres perceber que há solução para tudo. Obrigada por fazeres esta caminhada comigo.

Ameaça

Dizem que as pessoas mais inseguras são aquelas que mais projetam as suas inseguranças nos outros. Eu não tenho culpa das tuas inseguranças e do facto da tua vida não ter sido fácil. Por isso, vê lá o que fazes a mim e à minha família, porque eu sei que falas mal de nós pelas costas só pelo prazer de falar mal. Nós, que nunca te quisemos mal, que te ajudámos, que te apoiámos sempre. Sei também que pedes dinheiro ao teu pai, pela calada, sem dizer nada a ninguém, quando ele tem mais duas filhas e não dá dinheiro a nenhuma delas. És a única que ele ajuda porque és a "coitadinha", quando não és coitada nenhuma e só te aproveitas da boa vontade dele. Basta eu contar-lhe o que tu andas a dizer sobre mim e sobre a minha família, que ele tanto adora, para essa fonte de rendimento, que tanto te dá jeito, parar. Continua a falar mal de nós e a contar mentiras, que o teu pai vai ficar a saber a cobra que és, o mal que tens feito, e o teu dinheiro extra vai parar de aparecer. Para além do dinheiro, não te esqueças que tenho os contactos de alguns dos teus amigos e posso ligar-lhes a contar a mentirosa que és, intriguista e falsa que não merece a amizade de ninguém. Não percebo os teus motivos nem quero perceber. Já me fizeste mal a mim o suficiente para me preocupar com aquilo que te possa acontecer. Fizeste-me mal e vais sofrer as consequências. Vais pagar por aquilo que me tens feito e vais ter que começar do início para voltar a fazer amizades e para conquistar a confiança das pessoas que te tratavam bem. Vais provar do teu próprio veneno e sentir na pele o que custam as mentiras e as intrigas.

Extorsão

Caro vizinho, espero que o negócio do seu restaurante e pastelaria vá de vento em popa. Sei que tem tido bastantes clientes, que vende coisas para fora e que o lucro tem sido bastante. Tanto para eu poder fazer a seguinte proposta: ou me dá parte dos seus lucros ou eu divulgo o segredo que esconde de todos. A maior parte dos seus amigos e dos seus clientes pensa que é a pessoa mais afável e simpática do mundo, mas enquanto vizinha sei o que esconde e os negócios escabrosos que faz. Mas estou disposta a relevar se me der parte do lucro que obtém com o seu restaurante. Se me disser que não, estou disposta a vir revelar o seu "esqueleto no armário" por todos os meios que conseguir. Vou à rádio, aos jornais, à televisão, onde quer que seja para destruir completamente a sua reputação, o seu negócio e para mostrar que o lugar dos criminosos é na prisão. Se quer evitar este desvio no seu percurso aparentemente tão regular e feliz, aconselho-o a seguir aquilo que eu lhe digo. Caso contrário irei à polícia e aos meios de comunicação revelar como tem obtido dinheiro extra em negócios obscuros e ilícitos. É algo abominável e condenável, e tenho a certeza que todos lhe apontarão o dedo e o acharão asqueroso. A personalidade simpática e amorosa irá desvanecer e a pessoa verdadeira que é vai vir a público. Quero que me dê cinquenta por cento dos lucros do negócio do seu café em troca do meu silêncio. Basta sequer tentar renegociar comigo e vai tudo por água abaixo. Não aceito menos do que esse valor. É o preço que tem a pagar para que o seu segredo permaneça como tal. E se alguma vez eu perceber que me está a tentar trapacear, se a quantia de dinheiro for mais baixa do que aquilo que deveria ser, pode crer que o que esconde vai vir para a praça pública.

Reclamação

Gostaria de apresentar a minha reclamação quanto à forma como o meu processo de matrícula foi efectuado durante o ano lectivo de 2014/2015. É inadmissível que as funcionárias dos serviços académicos tenham perdido o meu processo de candidatura, uma vez que foi efectuado dentro dos prazos e seguindo os trâmites normais. Apresento também o meu profundo desagrado com a forma como trataram do meu caso, adiando a sua solução sempre o mais possível. Parece impossível que uma instituição como esta funcione de forma tão lenta e passiva, deixando que os problemas se arrastem em vez de os resolverem o mais rápido possível. As funcionárias da secretaria pareciam não saber o que fazer, fui eu que tive que insistir para que o meu processo fosse encontrado e tudo se resolvesse. Todos os dias ia à secretaria ou telefonava para os serviços académicos para saber se o meu caso já tinha sido resolvido, embora tudo parecesse mal parado. Tudo isto decorreu durante três semanas, quando podia ter sido resolvido em menos tempo. Só quando, numa conversa telefónica, comecei a gritar e a ameaçar com uma reclamação é que resolveram todo o meu problema. No próprio dia e numa questão de minutos. Acha admissível? Uma instituição tão prezada como é esta escola, com tão bom nome, ter funcionários de extrema incompetência e que demoram imenso tempo a resolver coisas aparentemente tão simples? Foi graças a duas professoras minhas que consegui resolver parte do meu problema e foram elas que encontraram o meu processo, aparentemente, perdido, quando isto era função das pessoas que trabalham na secretaria. Manifesto o meu profundo desagrado e descontentamento ao perceber que não posso confiar nos funcionários desta escola ou na eficiência dos seus serviços. Espero que esta reclamação surta algum efeito e sirva tanto para chamar a atenção para este tipo de questões, como para ajudar a melhorar a qualidade dos serviços prestados.

Informante FA

Agradecimento

Ver as estrelas sempre foi uma actividade favorita para mim. Toda a imensidão do espaço, ali, a olhar de volta para nós, tão perto e tão longe. O que haverá naquele espaço? Que outros mundos e vivências existirão no cosmos? Que planetas, paisagens, sóis, que outras criaturas existirão por aí?

Estas perguntas sempre flutuaram na minha mente. Ah, e como seria ser uma dessas pessoas, pioneiros do espaço que exploram locais onde (pensamos!) nunca ter passado um ser humano? Hoje posso dizer que sei como é ser um pioneiro, pois hoje estive no espaço pela primeira vez.

Este sonho persegue-me desde a infância, e eu acompanho-o, e alimento-o. Porque no espaço espera-nos todo uma nova existência que não temos noção que existe. Mas eu agora tenho. Finalmente comecei a cumprir o meu sonho. E devo-vos isso.

Deste meu ponto de visão privilegiado, estou mais perto das estrelas e de todo o espaço que espero um dia explorar. E tudo começou com vocês, que acreditaram em mim este tempo todo, que me aturaram e me apoiaram nesta minha loucura de fugir à Terra.

O que poderia ter feito sem pais que me acompanharam em noites de Lua nova a olhar para as estrelas, que me ensinaram os nomes das constelações e onde procurar os planetas visíveis? Sem essa atenção nunca poderia ter vindo a desenvolver tal interesse e paixão. Não tenho palavras para descrever quão agradecido estou, apenas que onde quer esteja, em que planeta esteja, ou em que local do universo esteja, vocês também estarão lá, comigo. Obrigado.

Foram os meus pais que alimentaram esta ideia de ver a Terra de longe, mas nunca teria conseguido sem o apoio dos meus amigos, que me aturaram dias e noites infindáveis a dizer nomes estranhos, a relatar notícias sobre calhaus cósmicos de que eles nunca quiseram saber e arrastá-los para noites frias para olhar só para o céu. Sem eles também nunca escreveria esta carta. Obrigado.

Ameaça

A paróquia é um sítio sagrado, onde todos nós gostamos de ir para reflectir em paz, no sossego do senhor. Falo por mim e acho que falo por todos que a frequentam quando digo que desejamos que esse local seja preservado e essa paz mantida. O local do Senhor é sagrado, e todos os que o frequentamos desejamos ser abençoados por essa luz divina. Para que esse local possa ser mantido, que continue a ser o recanto espiritual que todos gostamos, e para que possamos continuar a ajudar os mais necessitados são necessários bens materiais – dinheiro – já que nada deste mundo é grátis, tudo requer esforço ou recursos. Queremos que esse recanto do mundo continue a ser o nosso lugar de descanso. Por isso acho triste que a minha meditação e paz na paróquia seja interrompida, seja que por motivo for. O facto de o motivo ser roubo só aumenta a minha tristeza.

Tenho notado que das últimas vezes que agraciaste os restantes paroquianos com a tua presença, esta não tem mantido a paz que se espera naquele local sagrado. A caixa das doações acaba ligeiramente mais pobre após a tua passagem por ela. Essas doações são dadas para o bem da paróquia e dos paroquianos; aquele dinheiro é colocado lá para bem da população geral, e não da tua em particular. Como tal, estás avisado a que esses roubos são para serem cessados de imediato, sem qualquer tipo de tolerância futura. Uma pessoa precisar de ajuda pontualmente é uma coisa; se precisa sempre de ajuda, essa ajuda procura-se pelos meios adequados.

Por isso, se não parares os roubos contínuos que andas a fazer às esmolas da paróquia, serás expulso desta e nunca mais poderás voltar. Este espaço do Senhor ser-te-à vedado e as tuas acções serão tornadas públicas, para que todos saibam o que andas a fazer à comunidade – as tuas acções não podem continuar impunes.

Extorsão

Eu tenho um problema. Tenho filhos, e quero dar-lhes uma vida melhor. Por muito que trabalhe tenho dificuldade em sustentá-los devidamente. Mas espero um dia vir a ultrapassar essa dificuldade.

Ora, o senhor tem um certa visibilidade pública. Espero que a aprecie, deve ser interessante ser reconhecido quando se anda na rua. Ter pessoas que o acarinham e o seguem na sua vida.

Mas todas as medalhas têm o seu reverso.
Exactamente por ser conhecido, é que o reconheci há uns dias, na companhia de uma senhora que não era a sua esposa. Ainda para mais, o senhor acompanhou essa senhora a um motel onde os quartos são alugados à hora. Já no passado o senhor tinha sido vítima de um escândalo semelhante, tendo sido ilibado de tais acusações; sua esposa sempre confirmou que o senhor é fiel, e que nunca haveria terceiras rodas no vosso casamento. Dado isto, parece estranho a sua presença no descrito lugar como uma senhora que – perdoe o fácil julgamento – deve alugar o seu "amor" à hora.

Eu sou uma pessoa que acredita que há oportunidades no mundo, e que devemos todos beneficiar com elas. Neste caso, é oportuno eu o ter visto e fotografado com uma senhora que não é a sua esposa a entrar em tal sítio; se fosse uma escapadinha à rotina do casamento seria completamente compreensível. Mas isto tem todo o ar de ser uma escapadinha ao casamento. E acredito que a sua esposa irá achar o mesmo.

Como tal, acho que lucraríamos os dois fazendo um pequeno negócio. O senhor paga as despesas de educação dos meus filhos, e eu prontamente ignoro qualquer conhecimento ou prova que possua acerca do tão feliz encontro descrito acima. Afinal de contas, a sua esposa parece ser muito feliz consigo, e não desejo arruinar a felicidade de ninguém.

Será um prazer fazer negócios consigo.

Reclamação

Venho por este meio informar que os vossos produtos chegam danificados ao destino. Eu encomendei um dispositivo XPTO novo a partir do vosso site, e pedi entrega directa em minha casa. No dia da entrega encontrei uma caixa à porta de minha casa com a vossa identificação. Inspeccionando a caixa são evidentes as marcas demonstrativas de mau transporte, estando a caixa visivelmente amassada.
Abrindo a caixa foi é visível que o revestimento que protege o produto é insuficiente; este revestimento vinha destruído, não oferecendo qualquer tipo de protecção ao produto, permitindo que quaisquer pancadas ou danos que aconteçam sobre a caixa aconteçam também ao produto.

Retirando o produto da caixa, são visíveis as marcas deixadas pela falta de cuidados durante o seu transporte. É inadmissível que um produto novo venha cheio de covas, riscos e peças soltas, como é este o caso.
Apesar de tudo fui verificar se o produto funciona, apesar de certas peças virem soltas. As luzes acendem mas mais nada acontece. Experimentei utilizar o produto mas ele não responde, nem produz qualquer tipo de ruído. Dá a ideia os circuitos ligam mas o interior está danificado ao ponto de não responder. Resumindo, é um pisa-papéis caro.

Todas estes danos são visíveis nas fotos que envio em anexo, para que não tenham dúvidas deste relato. Além disso posso devolver-vos o produto, para que confirmem que não funciona.

Aconselho-vos a que se quiserem manter os vossos clientes ou angariar novos tenham mais cuidado na distribuição dos vossos produtos, e que chamem à atenção que os transporta desta maneira; é inadmissível entregar um produto não funcional novo a um cliente. Neste momento não posso recomendar os vossos produtos ou serviços, visto que não pude testar o produto e o serviço foi atroz.

Espero uma resposta rápida e que isto tenha sido apenas uma questão pontual.
Cumprimentos

Informante JC

Agradecimento

Querida Senhora,

Escrevo-lhe por não conseguir apresentar-lhe frente a frente um agradecimento, um dos mais sentidos agradecimentos que alguma vez farei, e porque o tempo urge. Na verdade, no que toca à gratidão, o tempo sempre urge e a palavra dita soa a vento passageiro, incapaz que é de deixar em si uma marca comparável à que as suas palavras e os seus gestos deixaram em mim.

Aliás, dizer que deixou algo em mim é muito pouco: deixar pode ser efeito de uma distracção, de um descuido, de um acto involuntário; mas, no seu caso, nada houve de desleixo, de irresponsabilidade ou de acaso; pelo contrário, tudo o fez por mim, ou tudo o que eu me fiz por meio de si, foi obra da sua atenção, do seu desvelo, do seu amor sem descanso, da sua entrega em cada dia, gratuita, absolutamente livre e total. Se de outras pessoas cuidou, se outros ajudou – e bem sabemos que sim! –, nunca isso se notou, nunca a sua atenção a outra pessoa significou uma desatenção para comigo, imerso que fui nessa sua capacidade infinda de amor. Assim, não é que tenha deixado algo em mim, fosse pelo desleixo que referi, fosse por vontade de impor certas características ou opções; não, deixou-me a mim em si, o meu coração no seu coração, mesmo quando as circunstâncias ditavam um certo grau de separação; deixou-me a mim em si, a minha alma almejando ser tão inteira como a sua, querendo fazer da minha vida um lugar de bondade e beleza como sempre vi ser a sua.

- Exagero! – responderia sem demora, se estivéssemos falando frente a frente, e pensará repetidamente, deveras envergonhada, quando ler estas insuficientes linhas. Mas não, o exagero foi seu, ao inspirar em mim a fuga à mediania, a possível grandeza de ser pessoa. Possível, sim, e, por isso, exigindo constante empenho, não em tarefas, não em coisas, não em ganhos, mas em ser sempre melhor, porque, como fez questão de me lembrar repetidamente, recuperando antiquíssimo e santo ensinamento, quando se deixa de querer ser melhor, deixa-se de ser bom.

Que eu, deixado em si, como quem no coração de outro se encontra a si mesmo, nunca deixe de ser bom, a única forma de ser verdadeiramente. Só assim a minha gratidão será consumada, só assim estas palavras serão verdadeiramente uma carta que alcança o seu escopo.

Devotadamente grato e verdadeiramente seu,

Ameaça

Ó seu grandíssimo escroque,

Não tem vergonha das patifarias que anda fazer?! De onde surgiu tanto ódio, para agora andar a prejudicar com mentiras e manobras sinuosas a minha vida e a das pessoas que me são mais próximas?

Eu procurei resolver a situação a bem, antes que fosse longe de mais, mas você insistiu e tem ultrapassado os limites do respeito e da decência, tornando isto quase num caso de polícia, portanto, digo-lhe agora muito seriamente: ou pára de agir como tem agido e procura emendar o mal que provocou, ou eu não só denunciarei explicitamente esta situação junto de quem de direito, como pedirei ao meu advogado para o processar por difamação e pelos danos causados.

Não creia que lhe escrevo de ânimo leve! Tenho sido paciente, mas o senhor já foi longe demais, pelo que lhe dou até ao fim desta semana para fazer o que lhe exijo. Se tal não acontecer, espero que esteja bem consciente de que o mal que lhe posso causar simplesmente ao tornar conhecida a situação será muito superior ao que me tem feito. A sua posição é já periclitante há muito tempo e, por isso, bastará uma palavra minha para acabar com o pouco de bom que ainda lhe sobra. E para tal bastará apenas que eu diga a verdade, nem sequer precisarei de descer ao seu nível ordinário e perverso.

Bem sei que a vida não lhe correu como queria, mas isso em nada se deve a mim ou às pessoas que me são próximas. Aliás, a sua permanente hostilização acabou por lhe destruir a única possibilidade que o senhor teria de fazer algo bom da sua vida. Portanto, este é o último aviso: corrija a sua atitude e os males causados, e talvez assim consiga algo de bom; caso contrário, será a sua vida, e não a minha, que sofrerá maiores danos a curto prazo.

Extorsão

Ex.mo Senhor

Quando viu que era eu quem lhe escrevia a presente carta, perguntou-se certamente qual o propósito da mesma, dado não termos contactos regulares, nem sequer uma relação de proximidade. No entanto, como perceberá, o motivo justifica não só este contacto, como que eu lhe tenha feito chegar esta carta sem que a mesma passasse pelas mãos da sua secretária. Actuei, afinal, visando o seu interesse: se alguém a lesse, o senhor ficaria numa situação melindrosa que poderia destruir num instante a sua carreira solidamente construída ao longo de tantos anos – assim a considera quase toda a gente, mas eu agora sei-a de uma debilidade facilmente denunciável.

Já terá percebido, ou, pelo menos, já teme, o assunto desta carta. É verdade, eu sei dos negócios pouco claros, digamos assim, que tem feito e que lhe têm permitido sustentar uma carreira aparentemente imaculada. Eu sei dos seus hábitos de quarta-feira à noite, onde se realizam e com quem. Como poderá ver pela amostra que junto a esta carta, não estou a fazer *bluff*: trata-se apenas de um exemplar de um vasto conjunto de provas que tenho reunido desde há algum tempo e que, obviamente, já reproduzi, estando devidamente guardadas e prontas a ser usadas, caso não tenha a gentileza de me julgar um seu comparsa, e mesmo quase amigo, que guardará estas provas com todo o cuidado, evitando que caiam nas mãos erradas – da polícia, por exemplo, ou daquela empresa que há tanto tempo o senhor tenta vencer. E isto, esta salvaguarda da sua vida como homem rico e poderoso, por apenas uma soma simpática de dinheiro. Considere-o uma prestação de serviços: eu guardo estas informações com todo o cuidado, e o senhor paga-me uma merecida quantia de dois milhões de euros.

Pergunta-se, certamente, tendo aprendido as devidas lições cinematográficas: "Mas que garantias tenho de que o seu silêncio está para sempre garantido ao dar-lhe este dinheiro?" Nenhumas, digo-lhe eu com a sinceridade que é própria de amigos que partilham segredos. E afianço-lhe desde já: este valor é apenas um primeiro presente da sua parte; dentro de uns tempos, precisarei de algo mais, dado que, como bem sabe, a vida não está fácil para quem não é um exemplar e imaculado homem de negócios como o senhor.

Em breve, dar-lhe-ei indicações práticas para que possa concretizar a sua oferta, selando assim a nossa amizade nascente. Até lá, sugiro que não defraude de forma alguma esta estima que já lhe tenho.

Reclamação

Ex.mos Senhores,

É com algum desgosto que, depois de tantos anos de colaboração, me encontro na necessidade de me dirigir a V.as Ex.as para fazer notar a forma descuidada, e mesmo legalmente incumpridora, como tenho sido tratado nestes últimos tempos nas nossas relações laborais.

Lamento, desde já, que esta carta surja na sequência de várias tentativas minhas de entrar em contacto com V.as Ex.as para evidenciar a insustentabilidade da situação presente, tentativas que se têm quedado sempre sem resposta, seja na mudança de atitude e resolução dos problemas, seja mesmo no simples e cordial cuidado de dar uma qualquer resposta à minha situação, tendo antes V.as Ex.as optado continuadamente por ignorá-la e por protelar a sua eventual resolução.

Venho, pois, reiterar o meu descontentamento relativamente aos incumprimentos da Vossa parte nos últimos tempos, a saber: a falta do devido pagamento dos últimos quatro meses de trabalho, a súbita desafectação do gabinete e respectivo serviço de secretariado com que tenho trabalhado desde o início e, como já referi, a falta de uma explicação e de qualquer tipo de atenção em relação a tudo isto.

Atingido, assim, o limite que posso suportar, não só financeiramente, tendo em conta a falta de pagamentos, como no que diz respeito às condições e aos compromissos laborais, peço uma última vez a V.as Ex.as que seja regularizada a situação, mediante o pagamento dos valores em atraso, que já tive a oportunidade de esclarecer junto da tesouraria, bem como mediante um esclarecimento das razões que conduziram à situação presente e uma explanação clara de perspectivas de futuro imediato, para que eu possa equacionar se há ou não condições para dar continuidade ao contracto que presentemente rege as nossas relações laborais.

Esperando a melhor atenção de V.as Ex.as para este assunto, que espero seja resolvido com brevidade, apresento cumprimentos cordiais.

Informante JV

Agradecimento

Escrevo esta carta para agradecer a minha família por tudo o que fizeram por mim.

Gostava de agradecer aos meus pais por me darem uma educação digna. Por dedicarem tempo a brincarem comigo e com a minha irmã enquanto pequenas e por me oferecerem uma infância muito feliz. Agradeço todo o carinho, paciência e dedicação. Agradeço-lhes por todas as vezes em que me ouviram e que aconselharam. Estou muito grata pela oportunidade que me oferecerem estudos superiores que irão tornar o meu futuro muito melhor. Também me ofereceram a oportunidade de viajar e conhecer outras culturas e países. Agradeço por me apoiarem nos bons e nos maus momentos. Por me ajudarem a escolher os melhores caminhos.

Agradeço a dedicação não só dos meus pais, mas também dos meus tios que sempre estiveram prontos para me ajudar em todos os momentos. Foram uns segundos pais para mim, que também me proporcionaram uma infância muito feliz, com muito carinho e dedicação.

Agradeço também aos meus avós pela paciência durante a minha infância, pela dedicação e carinho.

Em especial, agradeço também à minha irmã, que está sempre pronta a ajudar-me em qualquer ocasião. Está sempre comigo nos bons e nos maus momentos. Somos muito próxima e não imagino a minha vida sem ela.

Queria muito agradecer aos familiares que apesar de viverem longe também contribuem para a minha felicidade. Estão sempre prontos a ajudar-me mesmo estando longe.

Finalmente, quero também agradecer ao meu namorado que tem tido muita paciência e me dedica grande parte do seu tempo. E pelo carinho, dedicação e amor que recebo dele.

Obrigada a todos vós que fazem de mim uma pessoa melhor e que contribuem para que seja uma pessoa muito feliz.

Ameaça

Já sei o que tens feito nos últimos dias. Sei que tens andado a espalhar mentiras sobre mim. Nota-se que tens muita imaginação para andares a inventar coisas tão ridículas. Estou muito zangada! Ainda não consegui perceber o que é que ganhas com isso. Não ganhas nada!!! Aliás, só tens a perder! Quando toda a gente a quem contaste mentiras a meu respeito souber que é tudo invenção tua, vão achar que és um parvo mentiroso e nem vão querer estar ao pé de ti com medo que lhes faças o mesmo. És um pobre coitado desocupado e como não tens nada para fazer resolver abrir a boca para dizer dispartes e mentiras. Porque é que não arranjas qualquer coisa para fazer em vez de andares a prejudicar os outros. Ainda não percebi porque é que me queres prejudicar. É porque tens inveja de mim ou é pura maldade? E não é só a mim que prejudicas, também prejudicas a minha família, apesar de eles não acreditarem nas tuas mentiras. E os meus amigos também não acreditaram nas tuas mentiras, aliás foram eles que me contaram o que andaste a fazer. Só as pessoas que não me conhecem bem ou que não são inteligentes é que acreditam nas tuas mentiras sem fundamento nenhum. Não tens o direito de falar de mim ou de qualquer outra pessoa dessa forma. Não me conheces suficientemente bem para dizer o que quer que seja a meu respeito.

Se não parares com as calúnias, eu vou fazer queixa de ti. O que andas a fazer é difamação. Estás a por em questão a minha moral e integridade. Por isso pensa bem. Pensa duas vezes antes de falar. Se uma queixa não for suficiente pode ser que faça mais qualquer coisa, para ver se ficas de boca fechada e com a imaginação menos fértil...

Extorsão

Descobri umas coisas sobre si que sei que deseja muito esconder de todos os colaboradores da sua empresa e da justiça. Se o que descobri for tornado público pode acabar com a sua carreira e não só. Descobri que tem desviado dinheiro da empresa. E melhor do que ter conhecimento desses factos é ter provas. Como pode ver pelas cópias que estão dentro do envelope, as provas que possuo podem trazer-lhe muitos problemas, não só a nível profissional como a nível pessoal. Não se preocupe porque tenho muitas cópias e por isso pode destruir essas se quiser.

Para além de ver a sua carreira destruída, também corre o risco de ir preso durante vários anos. E também corre o risco de ver a sua família e amigos contra si.

No entanto o meu silencio tem um preço, basta oferecer-me uma boa quantia da sua vasta fortuna para que estas provas desapareçam por completo. Ou será que prefere que toda a gente fique a saber da sua falta de honestidade e de carácter?

Consegue imaginar a sua vida sem o luxo com o qual viveu durante tantos anos?

Consegue imaginar a sua vida sem a sua carreira empresarial?

Consegue imaginar-se preso?

Consegue imaginar-se sem amigos e com a família contra si?

A quem prefere pagar? A mim ou a um advogado muito, mas mesmo muito bom que não lhe pode garantir a liberdade? Mesmo que o advogado consiga o milagre de evitar que seja preso, vai acabar por ficar sem a sua carreira, o seu dinheiro e claro, não acredito que a sua família e amigos compreendam o seu ponto de vista.

Até quanto está disposto a pagar por o meu silêncio? Estou a pensar receber uma boa parte daquilo que lucrou ao longo destes anos. Seja muito generoso e não me desiluda!

Reclamação

Venho por este meio reclamar a minha espera perlongada neste hospital. Dirigi-me às urgências deste hospital às 11 horas do dia 11 de Julho de 2011 devido a uma alergia na pele. Recebi uma pulseira verde que, segundo a triagem o hospital, indica que se trata de um caso com pouca gravidade. Só fui atendia por um médico por volta das 20 horas do mesmo dia. Estive cerca de 9 horas com a pele irritada à espera de ser atendida, esta espera tão longa acabou por piorar o meu estado de saúde. Quando fui atendida o médico fez um diagnostico rapidamente e receitou os medicamentos necessários. Não compreendo porque é que casos de fácil resolução tenham que ficar pendentes durante tanto tempo. Apenas foram necessários cinco minutos do tempo do médico, mas para isso tive que ficar 9 horas numa sala de espera com a pele irritada a piorar ao longo do tempo e sem qualquer apoio. Acho que nenhum hospital está preparado para ter utentes durante tanto tempo numa sala de espera, havia muito poucas condições. Por exemplo, apenas dispõem de uma máquina de snacks. De facto, não fui a única paciente que esperou tantas horas neste hospital naquele dia. Compreendo que haja pacientes com problemas mais graves e que por esse motivo tenham que ser atendidos com maior urgência. Mas acho inadmissível uma espera tão longa, qualquer que seja o caso do paciente. No meu caso em particular, acho ainda mais inadmissível, já que tive que permanecer no hospital tanto tempo sem nenhum medicamento para atuar rapidamente sobre a alergia que piorava ao longo do tempo. Gostaria que justificassem o motivo para uma espera tão longa. Espero que tenham em conta este caso e que atuem para evitar que os utentes tenham que esperar tanto tempo para serem atendidos.

Bibliografia

Adam, J-M. (1992) *Les textes: types et prototypes*. Paris: Nathan Université.

Almeida, D. (2014) Atribuição de autoria com propósitos forenses. *ReVEL– Revista Virtual de Estudos de Linguagem*. 12 (23). 148–186.

Barreto, F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M. F., Nunes, F., e Silva, J. R. (2006). Open Resources and Tools for the Shallow Processing of Portuguese: the TagShare project. *Proceedings of the V International Conference on Language Resources and Evaluation – LREC2006*. Genova, Italy.

Branco, A. & Silva, J. R. (2006). A suite of shallow processing tools for portuguese: Lx-suite. Em *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations* (179–182). Association for Computational Linguistics.

Castro, I. (2006). Norma linguística e ensino do português. *Caderno Escolar*, *Pensar a escola* (3), 30–34.

Chaski, C. E. (1997). Who Wrote It? Steps Toward a Science of Authorship Identification. *National Institute of Justice Journal.* 233 (233). 15–22.

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, *8* (1), 1–65.

Chaski, C. E. (2013). Best Practices and Admissibility of Forensic Author Attribution. *Journal of Lay and Policy*, *21* (2), 333-376

Chesire, J. (2002). Sex and Gender in Variationist Research. Em J. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change*. Malden MA: Blackwell Publishers.

Coulmas, F. (2005). *Sociolinguistics: The study of speakers' choice*. Cambridge University Press.

Coulthard, M. (2004). Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics*, *25*(4), 431–447.

Coulthard, M. (2006). ...and then... Language Description and Author Attribution. Disponível em: http://www.aston.ac.uk/lss/staff-directory/coulthardm/ (último acesso em maio de 2016)

Coulthard, M. & Johnson, A. (2007). *An introduction to forensic linguistics: language in evidence*. London; New York: Routledge.

Coulthard, M. & Johnson, A. (Eds.). (2010). *The Routledge handbook of forensic linguistics*. Milton Park, Abingdon, Oxon; New York, NY: Routledge.

Coulthard, M. (2013). On Admissible Linguistic Evidence. *Journal of Law and Policy*, *XXI*(2), 441.

Coupland, N. (2007). *Style: language variation and identity*. Cambridge, UK; New York: Cambridge University Press.

Coyotl-Morales, R. M., Villasenor-Pineda, L., Montes-y-Gomez, M. & Rosso, P. (2006). Authorship Attribution Using Word Sequences. *Lecture notes in computer science.*, (4225), 844–853.

Cunha, C. F. & Cintra, L. F. L. (1984). *Nova gramática do português contemporâneo*. Lisboa: Sá da Costa.

De Beaugrande, R. (1998) Language and Society: the real and the ideal in linguistics, sociolinguistics and corpus linguistics. *Journal of Sociolinguistics, (3) 1*, 128-139.

De Vel, O., Anderson, A., Corney, M., e Mohay, G. (2001). Mining Email Content for Author Identification Forensics. *Sigmod Record, 30,* 55-64.

Delgado-Martins, M. R. (1973). Análise acústica das vogais tónicas em Português. *Boletim de Filologia*, (22), 303–314.

Dicionário do Português Atual Houaiss. (2011). Lisboa: Círculo de Leitores e Sociedade Houaiss-Edições Culturais Lda.

Diederich, J., Kindermann, J., Leopold, E. & Paass, G. (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence*, *19*, 15.

Duarte, I (2003) A família das construções inacusativas. Em *Gramática da língua portuguesa* (5.ª ed.), 507-538. Lisboa: Caminho.

Faria, I. H., Ribeiro Pedro, E., & Duarte, I. (1996). *Introdução à linguística geral e portuguesa*. Lisboa: Caminho.

Fissette, M. (2010). *Author identification in short texts* (Bachelor Thesis (Dep. of Artificial Intelligence)). Radbound University, Nijmegen, The Netherlands.

Gibbons, J. (2003). *Forensic linguistics: an introduction to language in the justice system*. Malden, Mass.: Blackwell Pub.

Gibbons, J., & Turell, M. T. (2008). *Dimensions of forensic linguistics*. Amsterdam, NL: John Benjamins Pub.

Gillier, R. (2011). *O disfarce da voz em Fonética Forense* (Tese de Mestrado). Faculdade de Letras da Universidade de Lisboa.

Grant, T., & Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics*, *8* (1), 66–79.

Grant, T. D. (2010). "Text messaging forensics: Txt 4n6: idiolect free authorship analysis?" em Coulthard, M. e Johnson, A. (eds.) *Roultledge Handbook of Forensic Linguistics*. Routledge Handbooks in Applied Linguistics. London: Routledge. 508–522.

Hazen, K. (2002). The Family. Em J. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change*. Malden MA: Blackwell Publishers.

Hirst, G. & Feiguina, O. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, *22* (4), 405–417.

Johnson, A. e Wright, D. (2014). “Identifying idiolect in authorship attribution: an n-gram textbite approach” em *Language and Law / Linguagem e Direito, Vol. 1(1*). 37-69

Juola, P. (2006). Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1 (3), 233–334

Kenny, A. (1982). *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities.* Oxford: Pergamon Press.

Koppel,M., Argamon, S. & Shimoni, A.R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17 (4), 401–412.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, *60* (1), 9–26.

Kotzé, E. (2010). Author identification from opposing perspectives in forensic linguistics. *Southern African Linguistics and Applied Language Studies*, 28 (2), 185–197.

Labov, W. (1966). *The Social Stratification of English in New York City* (Second Edition: 2006) Cambridge: Cambridge University Press.

Litosseliti, L. (2010). Research Methods in Linguistics. London; New York: Continuum.

Lorena, A. C. & de Carvalho, A. C. (2007). Uma introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada*, *14*(2), 43–67.

Luyckx, K. e Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. *Proceedings of the ... International Conference on Computational Linguistics*, *1*, 513–520.

Mateus, M. H. M. e Cardeira, E. (2007) *Norma e variação.* Lisboa: Editorial Caminho.

Marquilhas, R., & Cardoso, A. (2011). O estilo do crime: A análise de texto em estilística forense. Em A. Costa, C. Flores, & N. Alexandre (Eds.), *XXVII Encontro Nacional da Associação*

Portuguesa de Linguística - Textos selecionados (pp. 416–436). Lisboa: Associação Portuguesa de Linguística.

Martins, F., Rodrigues, C. & Brissos, F. (2014). Fronteiras do vozeamento na identificação do falante. Em *Textos Selecionados*. Porto: APL.

Martins, F., Rodrigues, C., Brissos, F. & Simões, D. (2012). Parâmetros acústicos em perícias forenses na identificação do falante. Apresentado na 3rd European Conference IAFL, Universidade do Porto.

McCombe, N. (2002). *Methods of Author Identification* (B. A.). Trinity College, Dublin, Ireland.

McMenamin, G. (2001). Style markers in authorship studies. *The International Journal of Speech, Language and the Law*, *8* (2), 93–97.

McMenamin, G. (2002). *Forensic linguistics: advances in forensic stylistics*. Boca Raton, Fla.: CRC Press.

Mosteller, F. e Wallace, D. L. (1964*).* Inference and disputed authorship: The Federalist*.* Reading, Mass: Addison-Wesley.

Mouton, P. G. (2000). *Cómo hablan las mujeres* (2.ª edición). Madrid: Arco/Libros, S.L.

Oliveira, F. & Mendes, A. (2013). Modalidade. Em *Gramática do Português*, *I*, 623–672. Lisboa: Fundação Calouste Gulbenkian

Olsson, J. (2004). *Forensic linguistics: an introduction to language, crime, and the law*. London; New York: Continuum.

Olsson, J. (2008). *Forensic linguistics*. London; New York: Continuum.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825−2830.

Pérez, M. F. (2007). Discurso y Sexo. Comunicación, Seducción y Persuasión en el Discurso de las Mujeres. *Revista de investigación Lingüística, Universidad de Murcia*, *10*, 55–81.

Rodrigues, M. da C. C. (2005). *Contributos para a análise da linguagem jurídica e da interação verbal na sala de audiências*. Universidade de Coimbra.

Silva, J., Branco, A., Castro, S., & Reis, R. (2010). Out-of-the-box robust parsing of Portuguese. Em *Computational Processing of the Portuguese Language* (pp. 75–85). Springer.

Solan, L. M. (2013). Intuition versus Algorithm: The Case of Forensic Authorship Attribution. *Brooklyn Journal of Law and Policy*, *21*, (pp 551-576).

Sousa-Silva, R. (2013). *Detecting plagiarism in the forensic linguistics turn* (Ph.D.). Aston University.

Sousa-Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Maia, B., & Oliveira, E. (2011). «twazn me!!! ;(» Automatic Authorship Analysis of Micro-Blogging Messages. Em R. Muñoz, A. Montoyo, & E. Métais (Eds.), *Natural Language Processing and Information Systems* (pp. 161–168). Berlin/Heidelberg: Springer/Verlag.

Sousa-Silva, R., Sarmento, L., Grant, T., Oliveira, E., & Maia, B. (2010). Comparing Sentence-Level Features for Authorship Analysis in Portuguese. *Lecture notes in computer science*, (6001), 51–54.

Spassova, M. S. (2007). The Relevance of Inter and Intra Authorial Variation in Authorship Attribution. Some Findings on Syntactic Identification Markers. Apresentado na 8th Biennial Conference on Forensic Linguistics/Language and Law, University of Washington, Seattle.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538–556.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, *35*(2), 193–214.

Svartvik, J. (1968). *The Evans statements: a case for forensic linguistics*. Göteborg; Stockholm: Almquist & Wiksell.

Trudgill, P. (2000). *Sociolinguistics: an introduction to language and society*. Harmondsworth, Middlesex, England; New York, N.Y., U.S.A.: Penguin.

Turell, T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, *17*(2), 211–250.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.