TEXTO DIGITAL
Revista de Literatura, Linguística e Artes

A paternidade de *Madalena*: um estudo de caso de atribuição de autoria

The paternity of Madalena: a case of authorship attribution

Deise Joelen Tarouco de Freitas[a]

a Universidade Federal de Santa Catarina, Santa Catarina, Brasil - deisej@superig.com.br

Palavras-chave: Atribuição De Autoria. *Hyperbase*. Literatura Brasileira. *Madalena*. Machado de Assis. Estilometria. Conto. Estatística Textual. Moreira de Azevedo.

Resumo: Este é um estudo de caso de atribuição de autoria ao conto *Madalena*, publicado no periódico *A marmota*, no Rio de Janeiro, em 1859. A causa da controvérsia sobre a autoria começou pelo fato de Machado de Assis e Manuel Duarte Moreira de Azevedo, ambos colaboradores do jornal à época, utilizarem a mesma assinatura, M. de A., em seus textos. Nosso estudo é baseado na estatística textual e utiliza o programa *Hyperbase*. Adotamos uma metodologia que combina técnicas e cálculos a fim de abranger o maior número de variáveis focadas nas características estruturais do *corpus*. Após o estudo comparativo de 37 gráficos e análises em árvore, relativos a um *corpus* formado por cinco bases de dados, com contos e romances, constatamos que a maioria deles aponta para Moreira de Azevedo como o provável autor de *Madalena*.

Keywords: Authorship attribution. Stylometry. *Hyperbase*. Short Story. Brazilian Literature. Machado de Assis. *Madalena*. Moreira de Azevedo.

Abstract: This article belongs to the area of study of authorship attribution. We study the case of a short story published in a newspaper called A Marmota, in 1859, in Rio de Janeiro, Brazil. The signature of the text was M. de A., the same signature used by two brazilian writers that collaborated with A Marmota, at the same time. One of them was Machado de Assis, the most iconic and canonic Brazilian writer, and the other was an almost unknown writer called Manuel D. Moreira de Azevedo. In this work, we use the program Hyperbase that can turn the corpus in a big database. We analyze here the structural data by differents algorithms and arrive to an authentication.

Il est clair, en effet, que pour chaque individu,
chaque milieu, chaque époque,
chaque genre il exist des mots, des Figuras,
des constructions privilégiées,
comme les métaphores de Victor Hugo
ou les imparfaits de Flaubert (...)[1]

1 O OBJETO

Este artigo faz parte de um trabalho de atribuição de autoria de maior extensão. Aqui, em razão da limitação de espaço imposta pela natureza da publicação, escolhemos apresentar somente os dados mais significativos para possibilitar a compreensão e manter a coerência de nosso trabalho.

O objeto deste trabalho é o conto *Madalena*, publicado no jornal *A Marmota*, no Rio de Janeiro, em 1859, sob a assinatura *M. de A*. A dúvida quanto à identidade de seu autor tem origem na existência, à época, de dois escritores que costumavam colaborar com o jornal e que utilizavam, por vezes, a mesma assinatura em forma de abreviação. Um deles era Machado de Assis[2], autor canônico da literatura brasileira; e o outro é Moreira de Azevedo[3], escritor menos conhecido no meio literário e que notabilizara-se, sobretudo, por obras de temática histórica sobre a cidade do Rio de Janeiro.

Segundo o pesquisador Galante de Sousa, não há uma edição de *Madalena* publicada em livro registrada na Biblioteca Nacional que pudesse ser comparada ao texto publicado em *A Marmota*. Então, a dúvida vem desde 1955, quando Galante publicou a sua *Bibliografia de Machado de Assis*. Para nosso trabalho, a fonte do conto em estudo é o livro *Dispersos de Machado de Assis*, publicado em 1965 e organizado pelo pesquisador francês Jean-Michel Massa.

2 O CAMPO DA ATRIBUIÇÃO, TEORIA E MÉTODO

O campo da atribuição de autoria não se limita à Literatura. Antes, é inter e multidisciplinar, tanto na concepção dos algoritmos estatísticos e das ferramentas tecnológicas que viabilizarão

[1] GUIRAUD, Pierre. *Essais de stylistique*, p. 62.
[2] Joaquim Maria Machado de Assis (Rio de Janeiro, 1839-1908), autor de *Memórias Póstumas de Brás Cubas* e *Dom Casmurro*, entre outras.
[3] Manuel Duarte Moreira de Azevedo Assis (Rio de Janeiro, 1832-1903).

a parte técnica da investigação quanto nos seus usos, pois tal procedimento serve para o esclarecimento de casos de plágio, para autenticação de documentos e para investigações criminais (Linguística Forense), entre outros.

Há dois tipos de fontes que são básicos para o investigador: as externas e as internas ao texto. No caso da literatura, já partimos do texto escrito, o que elimina indícios como o sotaque ou a prosódia. Temos, então, o texto escrito, mas este não é um texto qualquer. Ao contrário da Linguística Forense, que trabalha com o texto direto do autor – até manuscrito, em alguns casos – , o investigador literário tem como objeto um texto controlado, quase de múltipla autoria, uma vez que passou por várias pessoas: revisores, editores etc. Enquanto o linguista forense pode se valer de erros e vícios de linguagem que apontam de forma mais evidente para uma particularidade estilística do autor, o investigador literário tem em mãos um texto limpo de qualquer afronta à língua escrita. Se, por um lado, a natureza de nosso objeto nos impõe algumas limitações, por outro, também podemos colher alguns indícios em elementos fora do texto, como a época, o público-alvo, o local de publicação e a natureza desta (livro, periódico etc.).

As questões de atribuição são complexas e, no curso da história do campo, vários métodos e algoritmos foram testados a fim de identificar os autores de diversas obras. Em geral, quanto à metodologia, a maior parte dos estudos de atribuição costuma filiar-se a três tipos: os apoiados no léxico (número de palavras, número de vocábulos, altas frequências etc.); os apoiados nos elementos infrassemânticos (o comprimento de palavras e frases, a proporção de palavras instrumentais, perfil das palavras de acordo com a posição dos fonemas que as compõem etc.); e, finalmente, o tipo apoiado nas categorias gramaticais (a distribuição de cada categoria ou seu conjunto).

Uma tendência natural neste tipo de investigação é começar pelo levantamento do léxico. Vimos que os primeiros estudos baseavam-se em compilar dicionários, listas de palavras e listas de concordância. Tratava-se de pesquisa de caráter mais descritivo. Logo vieram os estudos com foco na distinção, na marca autoral, no estilo. Um dos primeiros cálculos utilizados foi o de extensão/comprimento de palavra.

Luong e Mellet (2003) observam que, em busca da medida da distância entre dois textos, é sobre o léxico que a maioria dos estudos se apóia, seja pelo levantamento do número de

palavras (N+ total de ocorrências constitutivas da cadeia linear do texto) ou de vocábulos (V + total de formas diferentes), pelo fato de ser a palavra o parâmetro por excelência, fácil de coletar e "intuitivamente pertinente".

Temos, portanto, o primeiro problema relacionado ao uso do léxico como principal variável com a finalidade de distinção autoral: a temática está relacionada a ele de maneira determinante. Como afirmam Luong e Mellet (2003), "*les distances lexicales sont extrêmement sensibles au thème et même au genre des textes comparés*"[4]. Então, se temos dois autores que escolhem o mesmo assunto a tratar, é natural que eles compartilhem uma grande semelhança em seu vocabulário. Se as obras forem de mesmo gênero, a semelhança acentua-se mais ainda.

Enquanto a simples contagem de palavras é sensível ao tema e ao gênero, a análise quantitativa dos elementos ditos "infrassemânticos", como extensão de palavra, extensão de frase, proporção de palavras funcionais, entre outros, são considerados por Luong e Mellet menos sensíveis aos fenômenos contextuais e responsáveis por resultados mais estáveis. No entanto, se os gêneros comparados forem diferentes, perde-se uma parte de tal eficácia. A propósito de tal método, os autores reconhecem que ele dá resultados mais estáveis, enquanto "*défie toute tentative d'interprétation et ne permet que des classifications aveugles*[5]".

Em compensação, de acordo com os autores, o terceiro método se mantém portador de sentido, além "*d'être moins sensible au thème*[6]" seguindo a afirmação de Labbé et Labbé (2003) que "*les contributions, à la distance intertextuelle, des vocables classés en fonction de leurs catégories grammaticales sont assez souvent significativement différentes de la moyenne*[7]".

Em um ponto, os vários autores estudados convergem: não é possível detectar a autoria de uma obra baseando-se no estudo de apenas uma variável, como ficou provado pelos primeiros estudos com extensão de palavra, por exemplo. Segundo Hockey, não há regras rígidas sobre

[4] As distâncias lexicais são extremamente sensíveis ao tema e mesmo ao gênero dos textos comparados. (Tradução nossa).

[5] Desafia toda tentativa de interpretação e permite somente classificações às cegas. (Tradução nossa).

[6] De ser menos sensível ao tema. (Tradução nossa).

[7] As contribuições, à distância intertextual, dos vocábulos classificados em função de suas categorias gramaticais são muitas vezes significativamente diferentes da média. (Tradução nossa).

como proceder neste tipo de estudo, a não ser o fato de que quanto mais material empregado, melhor, ou seja, "mais textos, mais testes". O ideal é a combinação de múltiplas variáveis e técnicas de cálculo.

Por fim, a análise da estrutura permite muitas variações técnicas e metodológicas, e sua eficácia já foi comprovada em vários estudos. Consideramos, porém, que é mais coerente e produtiva para o pesquisador a opção de mesclar técnicas e variáveis a fim de ampliar ou aprofundar sua investigação.

Estilo é o conceito-chave que norteia os estudos de atribuição de autoria. Dubrocard e Luong (1999), evocando a obra *Problèmes et méthodes de la statistique linguistique,* de 1959, afirmam que Pierre Guiraud já desejava ver se constituir uma "*caractérologie du langage*" capaz de evidenciar traços característicos, próprios de um autor e/ou mesmo de uma obra, evidenciados por padrões de escolhas linguísticas, fossem elas de natureza consciente ou inconsciente por parte do autor.

Mas será que o estilo, em se tratando de literatura, é totalmente pessoal?

Segundo Anna Jaubert (1999), "*le style (...) se construit dans un mouvement de navette entre le pôle du particulier et celui de l'universel*[8]" Tal oscilação entre estes dois pólos nos remete ao problema da articulação entre os três aspectos essenciais que, conjuntamente, permitem a existência de uma obra literária: o gênero, a cronologia e o estilo pessoal do autor. Por isso, escolhemos estes três critérios como orientadores de nosso trabalho. Essa oscilação entre o particular e o universal pode evidenciar-se em vários aspectos. O que nos importa, primeiramente, é que Jaubert usa tal afirmação para alertar sobre como o gênero literário escolhido age como um precursor do estilo, na medida em que impõe as restrições estilísticas específicas de sua natureza, conforme nos indica a assertiva de Guiraud que serve de epígrafe a este trabalho.

Se tomarmos um *corpus* de mais de um autor, com obras de diferentes gêneros misturados, é por este critério que elas se agruparão em termos de distância lexical. Os contos com os

[8] O estilo constrói-se em um movimento de vaivém entre um pólo particular e outro universal. (Tradução nossa).

contos, os romances com os romances, o que prova a prevalência deste critério sobre o da autoria. Eis o primeiro fator que é determinante de certas características de uma obra.

Outro fator que não é pessoal é a época, a cronologia. Todo tempo tem suas características linguísticas próprias, o famoso *Zeitgeist* dos alemães, o espírito da época, muito mencionado nos estudos literários.

Os muitos estudos já feitos mostram que o fenômeno é complexo e que vários são os fatores que concorrem para a constituição da marca autoral e a consequente dificuldade de se isolarem as variáveis ou a combinação de variáveis que sejam exclusivas para tal identificação. É preciso garimpar com muita cautela o tesouro lingüístico/literário de um autor ou obra. Entretanto, mesmo depois de alguns séculos de pesquisas, ainda não há um consenso na área em relação a um método ou algoritmo que seja plenamente eficaz. O que não significa que não houve avanço nesse campo, visto que os acertos e, principalmente, os erros ajudaram a corrigir a direção dos estudos e apontaram para um conjunto das variáveis mais significativas e eficazes para a investigação de autoria, bem como das técnicas de abordagem e dos cálculos (algoritmos) mais acurados para se atingir um resultado válido.

3 O *CORPUS* E O MODELO DE ANÁLISE

Para construirmos o nosso modelo de análise, partimos do estudo do estado da arte do campo da atribuição de autoria. E, pelo histórico já apresentado, temos que o estudo do léxico é o ponto de partida natural, afinal a palavra é a matéria-prima do escritor.

No entanto, é preciso observar também a composição da obra, ou seja, a ordenação da sequência dos elementos textuais que consiste num "sistema ordenado de posições de fases em que uma fase se funda em fases correspondentes de todos os estratos conexos da obra e assim alcança determinadas qualificações precisamente por se encontrar nesta e não em outra posição", como ensina Ingarden (1965). Isto é, a organização faz parte da construção de sentido.

Nosso modelo de análise baseia-se em um *corpus* de textos digitalizados e no programa de estatística textual Hyperbase. Os principais critérios a orientar nossos procedimentos são

autoria, cronologia e gênero. O *corpus* foi lematizado, a fim de possibilitar a observação de fenômenos/dados relativos às estruturas sintáticas.

O material analisado contém a obra completa de Machado de Assis, cinco textos digitalizados de Moreira de Azevedo e o conto *Madalena*, objeto central desta investigação. Embora tenhamos constituído muitas bases diferentes para os testes desta pesquisa, na versão final do trabalho, usamos as seguintes bases de dados:

> **1) CONTOMAD** – é uma base de dados composta por 18 textos/obras que somam 433.843 ocorrências (N) e 25.723 vocábulos (V). Ela privilegia o gênero, apesar do caráter um tanto quanto híbrido dos textos de Moreira. Os textos foram colocados na base por ordem de autor e dentro desta sequência, em ordem cronológica, na exata ordem que aparece na lista a seguir: Ma1 - Três tesouros perdidos (58); Ma2 – Bagatela (59); Ma3 – O país das quimeras (62); Ma4 – Virginius (64); Ma5 – Questão de vaidade (64); Ma6 – Casada e viúva (64); Ma7– Cinco mulheres (64); Ma8 – O que são as moças (65); Ma9 – Uma excursão milagrosa (66), Ma10 – O oráculo (66); CF – Contos Fluminenses (70); HM – Histórias da meia-noite (73); PA – Papéis avulsos (82); HS – Histórias sem data (84); Mo1 – Madalena (60); Mo2 – Lourenço (68); Mo 4 – Homens do passado (75); Mo5 – No tempo do rei (89);
>
> **2) MOMAMAG** - composta por 18 textos/obras que somam 385.324 ocorrências (N) e 23.953 vocábulos (V), esta é uma base onde se misturam os gêneros, tendo conto e romance, na ordem da lista que segue: Ma1 – Três tesouros perdido (58); Ma2 – Bagatela (59); Ma3 – O país das quimeras (62); Ma4 – Virginius (64); Ma5 – Questão de vaidade (64); Ma6 - Casada e viúva (64); Ma7 – Cinco mulheres (64); Ma8 – O que são as moças (65); Ma9 – Uma excursão milagrosa (66); Ma10 – O oráculo (66); Ma11 – Ressurreição (72); Ma12 – Casa Velha (85); Ma13 – Quincas Borba (91); Mo1 – Madalena (60); Mo2 – Lourenço (68); Mo3 – Os franceses no RJ (70); Mo 4 – Homens do passado (75); Mo5 – No tempo do rei (89).

Usamos como ferramenta técnica e tecnológica a estatística textual auxiliada por computador. Trata-se, mais especificamente, do programa *Hyperbase*. Entretanto, adotamos um modelo híbrido de abordagem, aliando critérios qualitativos e quantitativos, na medida em que tanto o recorte, a escolha do *corpus* e a seleção das obras que compõem cada uma das bases de dados

utilizadas quanto o seu ordenamento foram norteados por critérios qualitativos, tais como autoria, cronologia e gênero literário. Além disso, nossa ferramenta dispõe de muitos algoritmos diferentes: desde meras listagens, como o dicionário gerado a partir do *corpus* e as listas de concordância às análises multidimensionais, como a análise fatorial e em árvore, até os mais complexos recursos, como as funções estatísticas amparadas nas estruturas sintáticas ou aqueles relativos às associações entre palavras (palavra pólo).

Optamos por buscar subgrupos ou subcategorias, a partir das informações de dados macro, ou seja, a partir da análise de categorias mais abrangentes, como as palavras funcionais, de modo a examiná-las individualmente. Por exemplo, depois de observar o comportamento do conjunto das funcionais, observamos o comportamento individual das conjunções, que formam um subgrupo do primeiro. E, dentro desta categoria, escolhemos alguns pares de sinônimos, para testar a tendência apontada nos passos anteriores.

Na sequência, listamos as categorias escolhidas para nosso estudo de caso e avaliamos o modo como elas foram combinadas e agrupadas, a fim de observar as variáveis de diferentes pontos de vista:

- Pessoas do discurso;
- Palavras funcionais a partir de listas;
- Categorias gramaticais – no conjunto e individualmente;
- Substantivos – teste com sinônimos;
- Verbos – teste com sinônimos;
- Tempo – teste com sinônimos;
- Preposições e demonstrativos;
- Demonstrativos;
- Preposições;
- Conjunções.

Por fim, consideramos alguns dados lexicais, a título de complemento.

Neste trabalho, como optamos por uma abordagem que combina técnicas e contempla muitas variáveis estruturais, resolvemos tomar apenas o que consideramos de mais essencial de cada uma, para depois, pela visão de conjunto, tentar chegar a alguma consideração bem fundamentada sobe o caso *Madalena*.

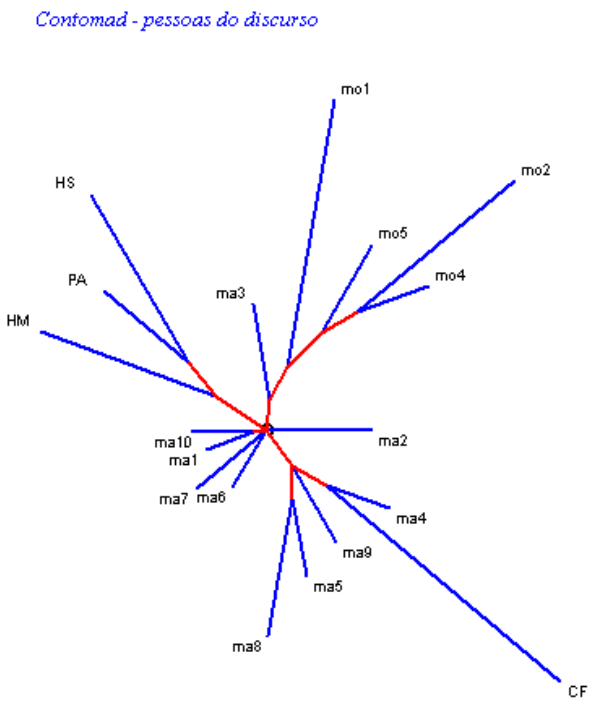
Na análise multidimensional em árvore, escolhemos representar as duas variáveis, coordenação e subordinação, pelo desvio, ou seja, por números relativos. Ao agrupar as variáveis, podemos notar uma mudança nas associações que aproximam os textos de Moreira de *Madalena*, ainda que ela também se mantenha relativamente próxima de ma3 e ma9. As folhas dos volumes continuam agrupadas por época e os avulsos dividem-se, ficando cinco para cada lado, e mantendo-se relativamente equidistantes entre si. De qualquer forma, *Madalena* mostra uma aproximação importante com Moreira.

O movimento de nossa análise vai do macro para o micro. Começamos pelo conjunto das categorias e passamos a particularizar cada uma, de acordo com a divisão dada pelo programa. Por isso, a partir de agora, passaremos à apresentação dos dados relativos a algumas categorias individuais, a fim de avaliarmos se o quadro de *Madalena* se altera.

3.1 PESSOAS DO DISCURSO

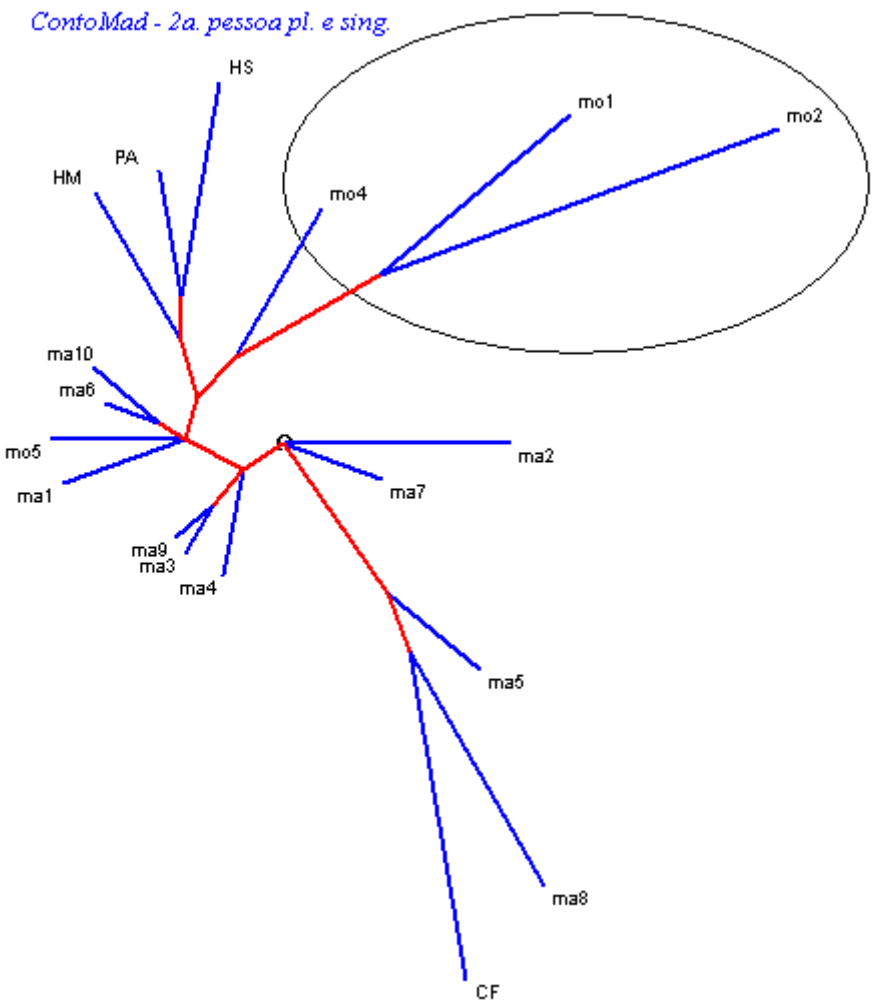
Verificaremos, a seguir, o comportamento dos textos a partir das pessoas do discurso. Comecemos pelo conjunto:

Figura 1: Contomad – pessoas do discurso.



A árvore acima é bastante interessante quanto aos agrupamentos, pois separa os avulsos (ma1 a ma10) em dois grupos distintos, tendo apenas ma2 isolado, enquanto, ma3, embora parcialmente isolado, aproxima *Madalena* do grupo de Moreira. CF destaca-se dos demais, apesar de estar próximo a quatro avulsos, em especial a ma4. Atribuímos este destaque do todo às longas passagens dialogadas do volume e à alternância de pessoa nessa configuração textual. Os três volumes restantes de Machado agrupam-se à esquerda, projetando um uso mais coeso das pessoas do discurso, aparentemente seguindo uma tendência cronológica, já que aparecem em ordem de publicação (no sentido anti-horário).

Figura 2. Contomad – 2ª. p. sing. e pl.

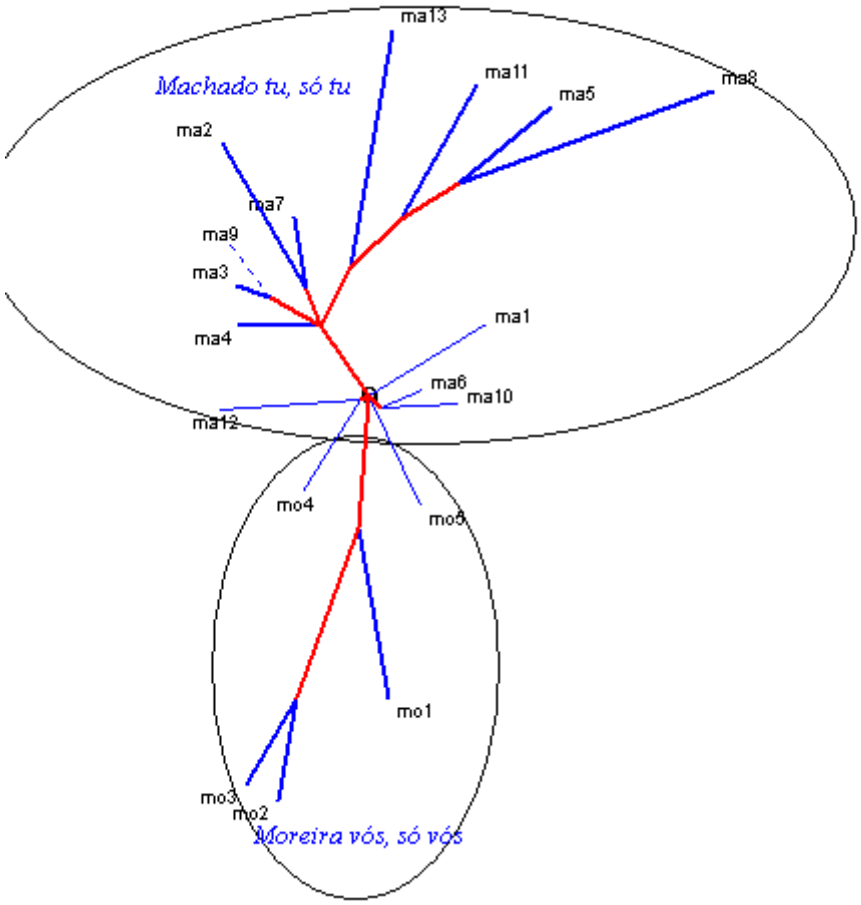


Embora semelhante à árvore das pessoas em conjunto, em relação ao distanciamento bastante significativo de CF e ao agrupamento dos outros volumes machadianos em oposição, esta figura não deixa dúvidas quanto à filiação de *Madalena* ao grupo de Moreira. Além de sair do mesmo nó que mo2, o ramo que habita só tem mais um vizinho, mo4, do mesmo autor. Os

avulsos continuam dispersos, o que se deve, supomos, ao fato de constituírem os primeiros passos de Machado no conto, ou seja, sua fase experimental nesse gênero.

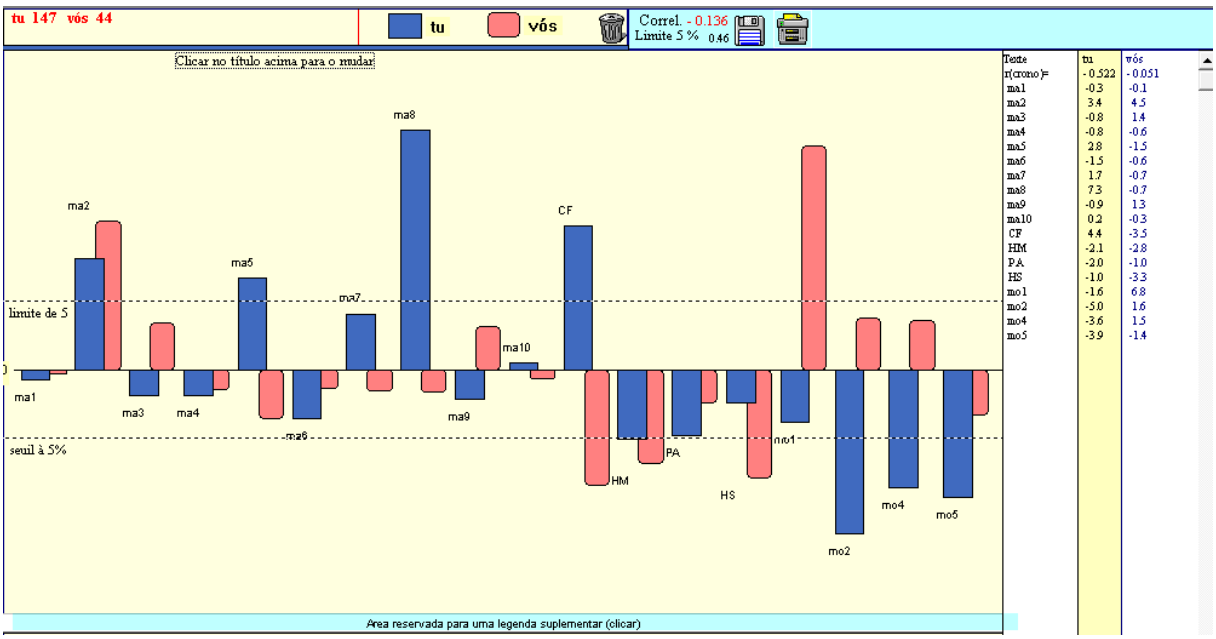
A seguir, observaremos o comportamento das variáveis em outra base, na qual mesclamos conto e romance:

Figura 3. Momamag – 2ª. p. sing. e pl.



É interessante observar que, ao misturarmos os gêneros, *Madalena* fica ainda mais próxima de Moreira. Outro dado que se destaca é a proximidade de mo4 e mo5, saindo do mesmo nó de três avulsos. A atração entre mo4 (Homens do passado, de 75) e mo5 (No tempo do rei, de 89) certamente não se dá pela via cronológica. A hipótese mais provável é que seja pelo gênero, pois mesmo que Moreira chame mo4 de "crônicas", sua estrutura é mais semelhante a um conto longo ou novela, como mo5.

Figura 4. Contomad - Tu e vós.



Mas em termos de pessoa, um diferencial que aproxima *Madalena* de Moreira de forma um pouco mais evidente é o uso da 2ª pessoa do plural. No gráfico acima, *Madalena* tem destaque por ter o maior superávit no uso de *vós*, expresso num desvio de 68, enquanto o segundo colocado é ma2.[9] *Madalena* é o destaque com o excedente mais significativo do *corpus* no uso do vós. Comparando-o com o Machado do início (avulsos) de quem costuma estar próximo, os únicos com quem tem afinidade quanto às duas variáveis são mo3 e mo9, que são praticamente o mesmo conto (m9 é reescrita de ma3). Portanto, quanto às duas variáveis, segue a tendência de Moreira, embora, nesta representação, tal proximidade não fique bem clara. Quanto ao uso do *vós*, talvez se torne mais clara a distância entre *Madalena* e Machado por meio da análise de um gráfico que traz a obra completa do autor, separada por gênero, conforme mostramos a seguir:

[9] *Bagatela*, de 1859, conto também controverso quanto à autoria. Melhor relativizar os dados a ele referentes.

3.2 PALAVRAS FUNCIONAIS

Figura 5. Contomad -Funcionais – ocorrências.

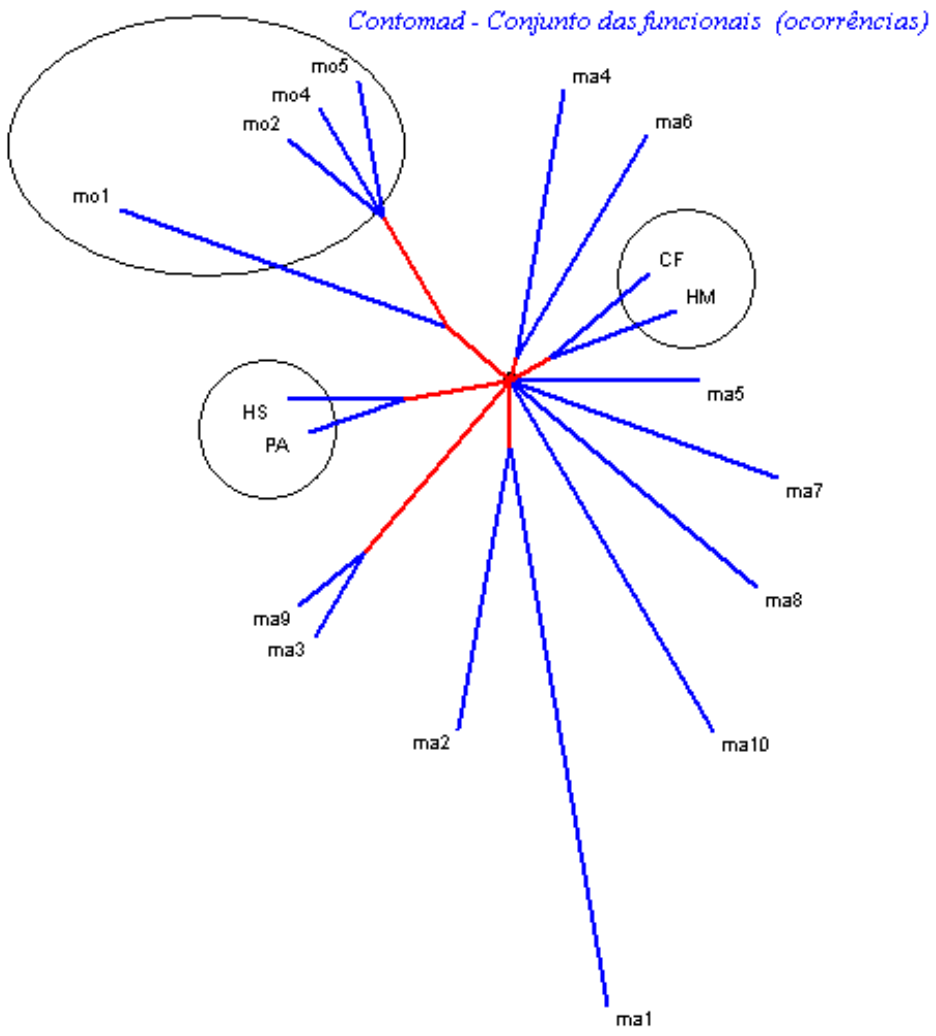
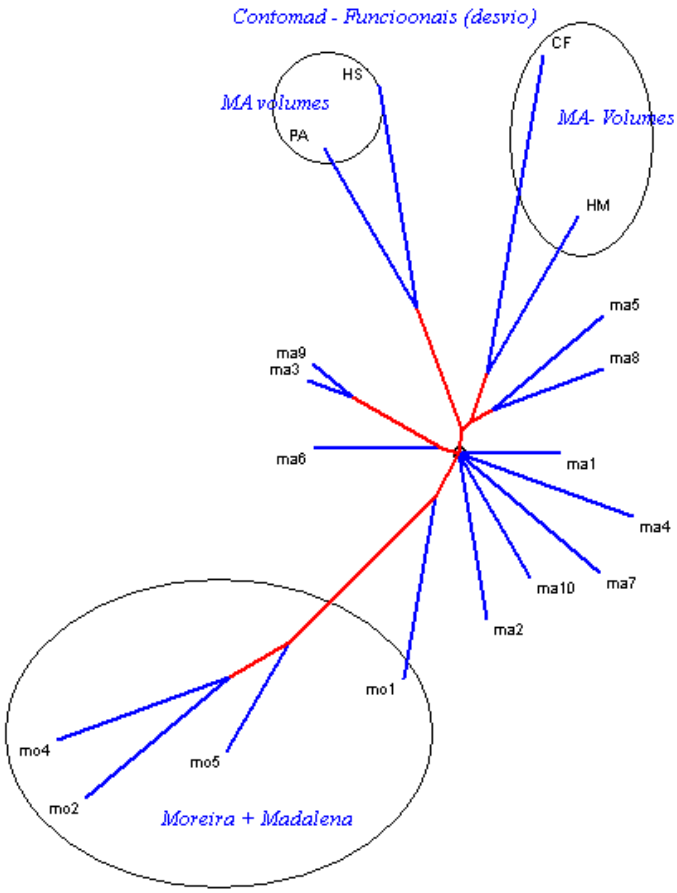


Figura 6. Contomad - Funcionais – desvio.

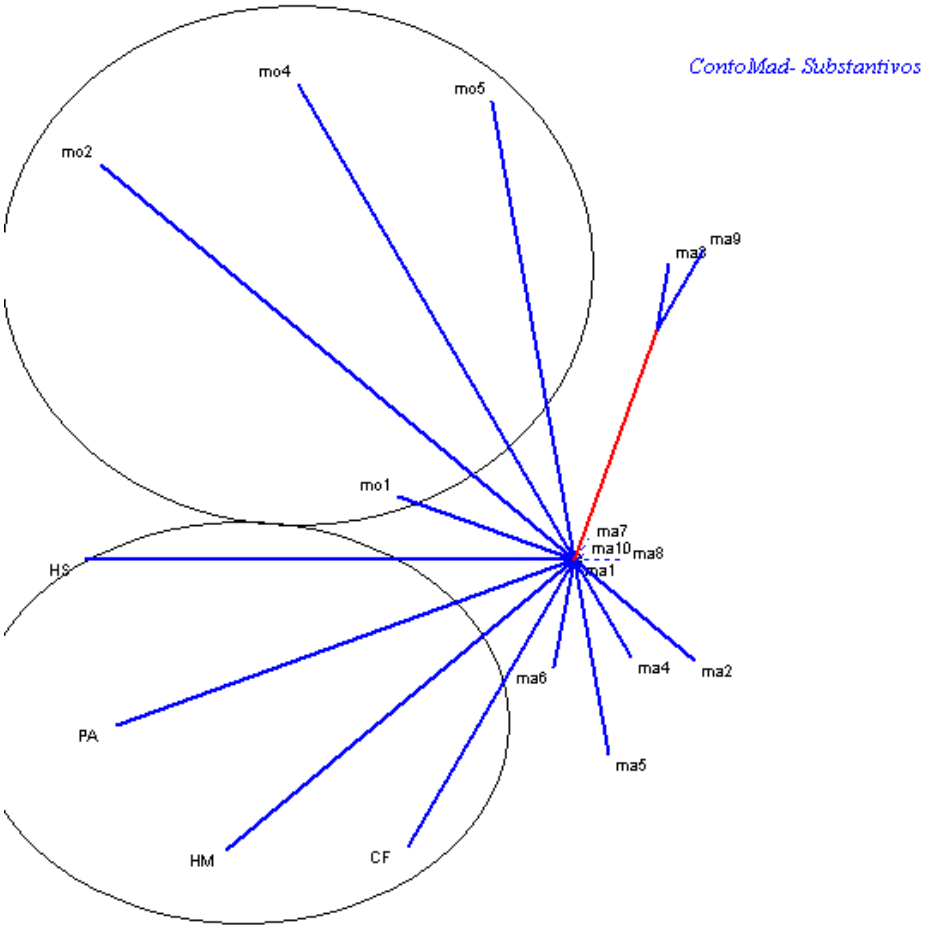


A árvore acima foi calculada por um conjunto de funcionais que transformamos em uma lista que compreende as palavras negativas, os pronomes demonstrativos, os artigos e as preposições. O conjunto das funcionais, diferente do conjunto das categorias gramaticais, coloca *Madalena* no grupo de Moreira. Seguindo os nomes e os verbos, a estrutura vai sedimentando a familiaridade entre eles. Como as funcionais fluem de forma mais automática, menos consciente, elas parecem, de fato, revelar as verdadeiras "relações" que não aparecem na superfície lexical.

Os volumes CF e HM e PA e HS, continuam agrupados, confirmando as semelhanças de época, ao mesmo em tempo que apontam para alterações estilísticas, pois os últimos se afastam dos primeiros. Se o uso das funcionais sofreu mudanças, foi alterada a maneira de estruturar o texto. Veremos mais alguns dados relativos a esse fator no decorrer deste trabalho.

3.3 SUBSTANTIVOS

Figura 7. Contomad – Substantivos – estrutura.



Esta árvore representa o conjunto dos substantivos, não a partir do léxico, de listas de nomes, mas das estruturas gramaticais formadas por este elemento textual, isto é, em que ele faz papel da constante. Assim, temos a oportunidade de visualizar como se dão os agrupamentos dos textos do *corpus* pelas combinações estruturais utilizadas.

O primeiro impacto dessa imagem é a quase perfeita equidistância entre a maior parte dos textos. Como já foi dito, m3 e m9 são muitíssimo semelhantes, devem ter dois ou três parágrafos diferentes apenas. Daí, a sua situação peculiar de distanciamento em relação aos outros. No entanto, a maior parte do *corpus* está separada pela conjunção gênero/cronologia.

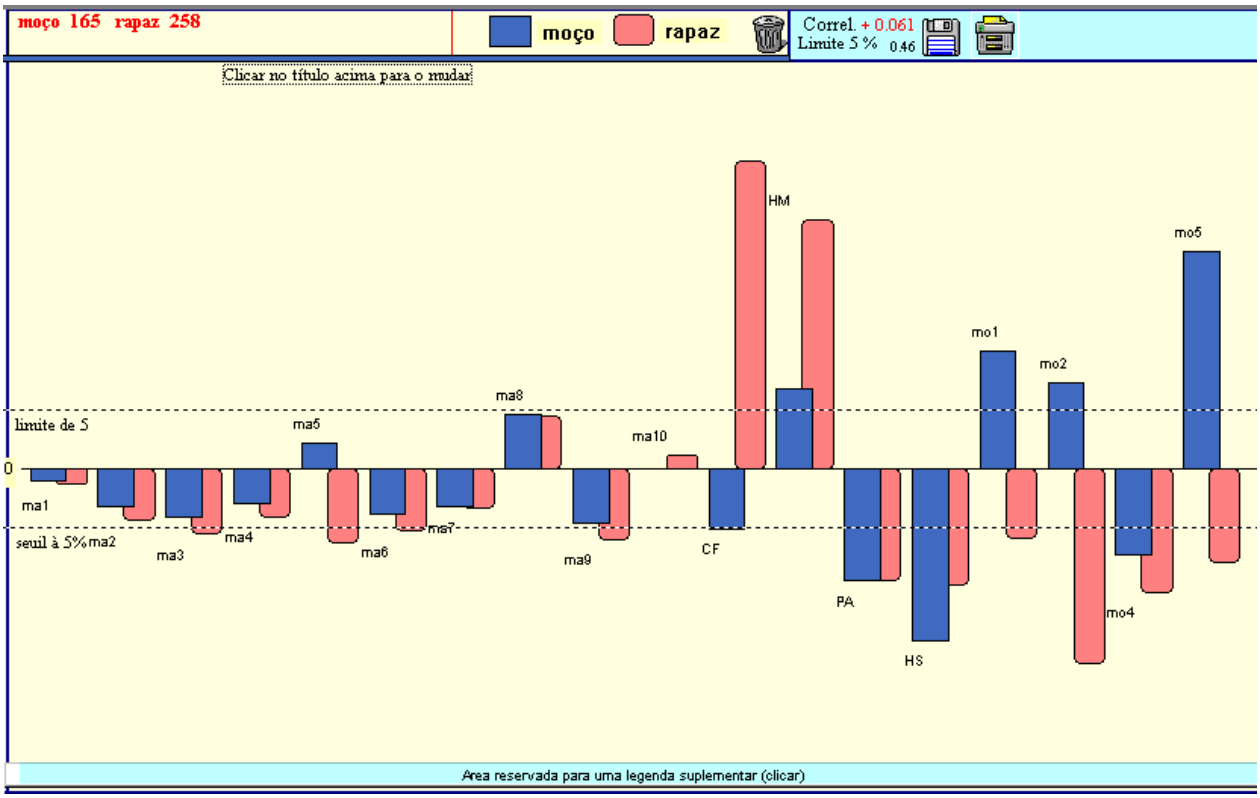
Em sentido horário, podemos começar com os avulsos de Machado, que partem do mesmo ponto, mas têm diferenças de comprimento significativas. Logo após, marcamos o grupo dos volumes de contos de Machado que aparecem em ordem cronológica. *Madalena* nos

possibilita dois tipos de leitura: podemos vê-la no meio do caminho entre os volumes de Machado e os textos de Moreira ou associada a Moreira, como aparece na marcação.

3.3.1 Substantivos – teste com par de sinônimos

Dado que a análise de pares de sinônimos mostrou-se um bom instrumento para fins de investigação de autoria, resolvemos testar alguns pares em determinadas categorias. Escolhemos os pares a partir das altas frequências e das palavras representativas do *corpus* disponíveis pelo cálculo de associações privilegiadas do *Hyperbase*. Para os substantivos, escolhemos o par composto de *moço* e *rapaz:*

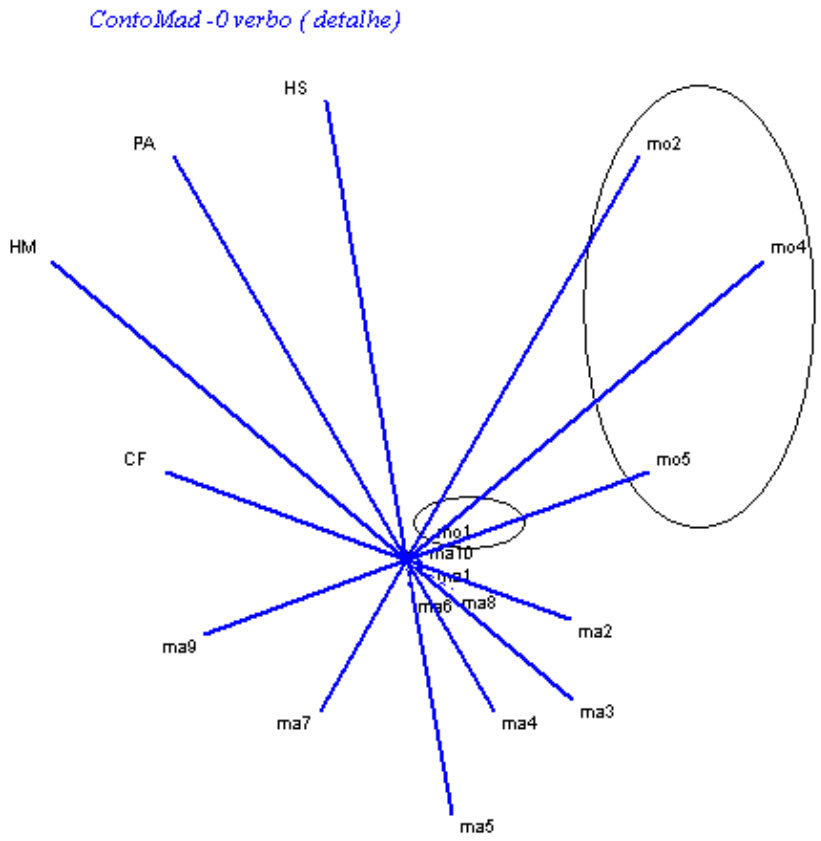
Figura 8. Contomad – moço e rapaz.



A palavra *moço* é uma das mais usadas na contística de Machado de Assis. Analisando o quadro acima, podemos afirmar que Machado apresenta dois altíssimos índices de *rapaz* em CF e HM, apesar de ter mais baixas do que altas no restante do grupo, no entanto, Moreira é uniforme nas baixos usos de *rapaz* e apresenta dois excedentes bastante significativos de *moço*, no que, mais uma vez, é acompanhado por *Madalena*.

3.4 VERBOS

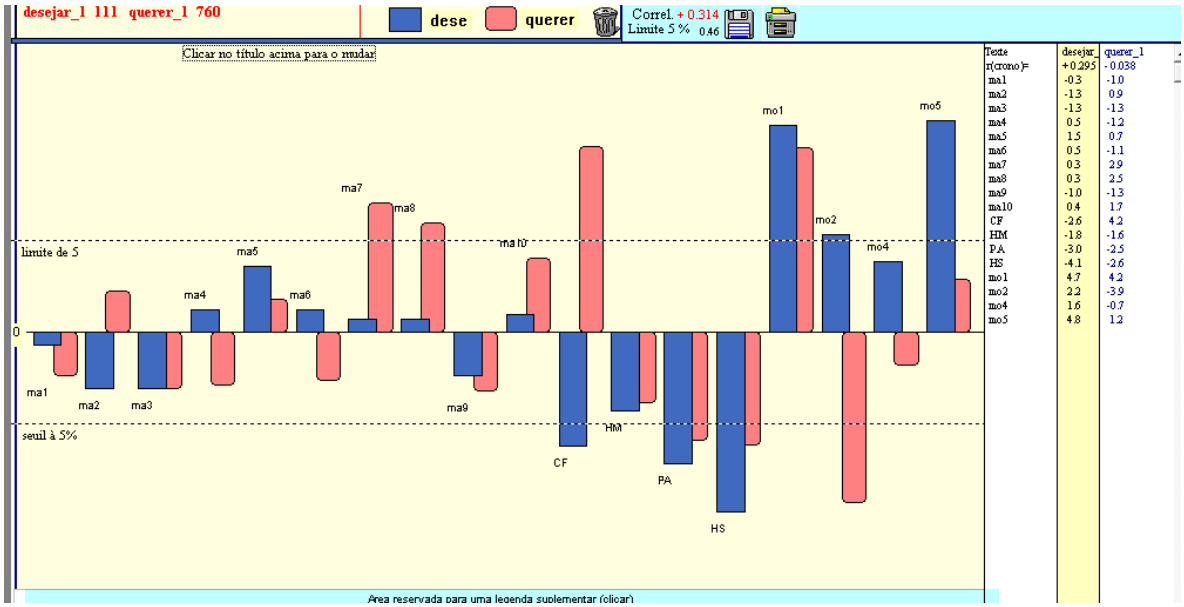
Figura 9. Contomad- verbo.



Partindo da estrutura, como no item anterior, temos, novamente, *Madalena* (mo1) agrupada a Moreira. Contudo, aqui temos um diferencial: ela está entre mo4 e mo5, ou seja, não há dúvidas de sua maior compatibilidade com Moreira, até porque a distância entre mo2 e HS é evidente, ao contrário da árvore anterior, onde os textos de Machado, *Madalena* e Moreira apareciam equidistantes. Em termos de estrutura, as semelhanças parecem ser bastante sólidas, com a tendência da árvore dos nomes sendo confirmada na dos verbos.

Como no item anterior, passamos ao teste com sinônimos dentro desta categoria:

Figura 10. Contomad- Querer e desejar.



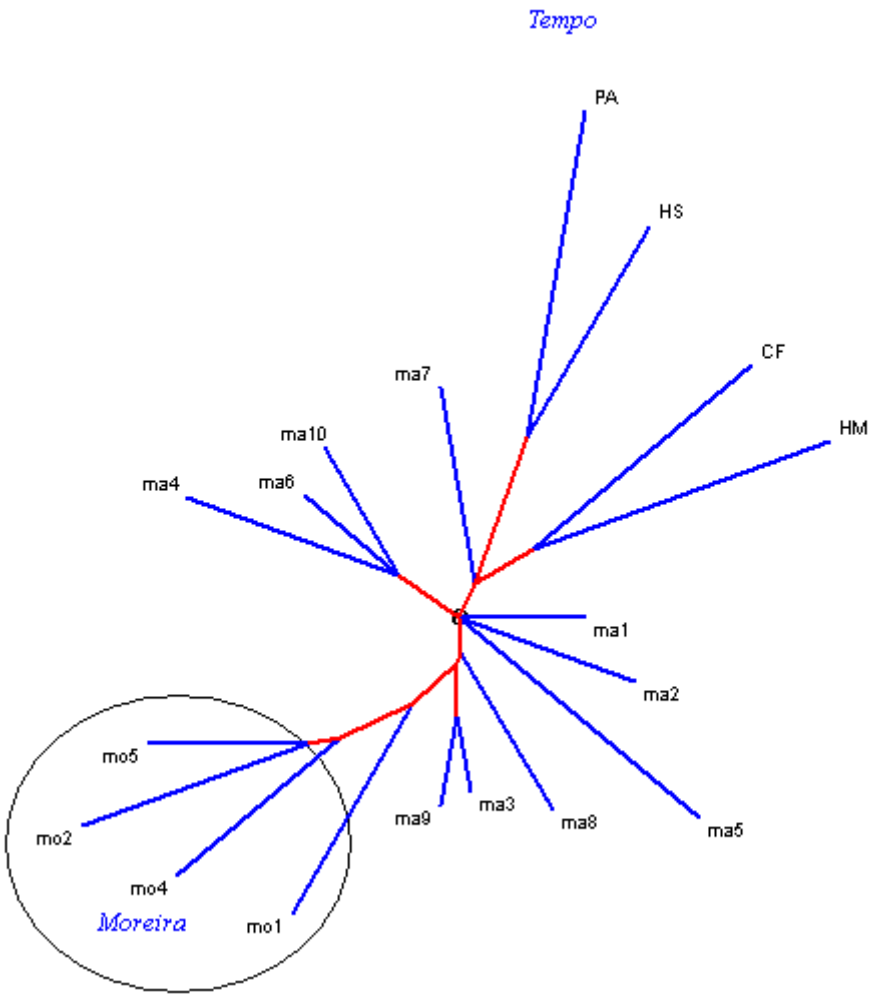
Eis *Madalena* e Moreira unidos novamente, e de maneira consistente, pois são os maiores excedentes do verbo *desejar* em todo o *corpus*. Machado apresenta 6 excedentes de *querer*, sendo 3 significativos do ponto de vista estatístico.

Continuamos, a seguir, com as demais categorias, por estrutura.

3.5 TEMPO

Outra categoria listada pelo *Hyperbase* é a de tempo, que traz os advérbios e conjunções desta natureza. Vejamos o que ela nos mostra:

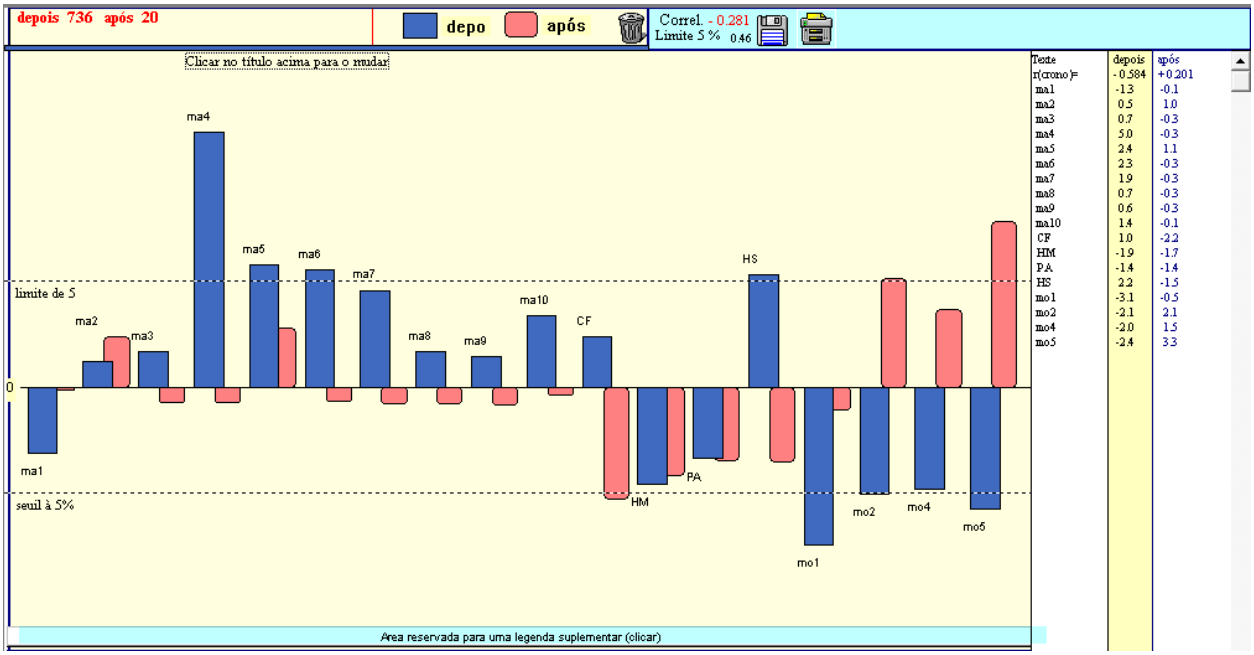
Figura 11. Contomad – Tempo (advérbios/ conjunções).



Mais uma análise em árvore que ratifica a filiação de *Madalena* ao grupo de Moreira. E não só isso: o grupo de Moreira se opõe aos volumes de contos de Machado. O que é significativo, pois, em outros dados vistos anteriormente, ficava clara a identificação de *Madalena* com os primeiros avulsos. Estes, embora tanto quanto dispersos, formam dois grupos, sendo que m3 e m9 ainda aparecem no mesmo ramo, relativamente próximos a *Madalena*.

Dentro da categoria de tempo, buscamos um par de sinônimos para testar:

Figura 12. Contomad – Depois e Após.

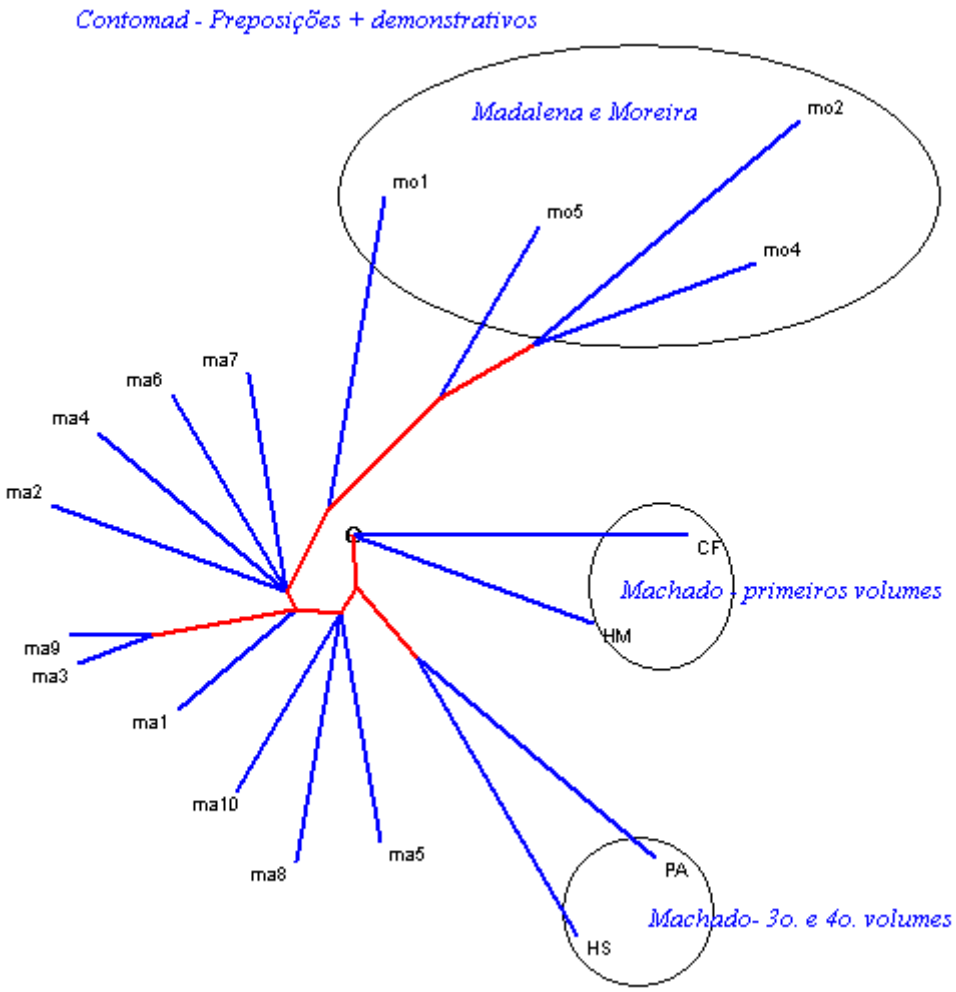


O gráfico relativo a tempo mostra-se bem mais definido em termos de divisão Moreira-Machado. O *depois* é claramente a primeira opção de Machado tendo apenas 3 dos 14 textos com baixas que não são estatisticamente significativas e 11 acima, sendo 4 significativas. Mas o que destaca a diferença quanto ao par são os baixos índices de uso de *depois* por Moreira e seu excedente na escolha de *após*. *Madalena* afasta-se de Machado, por sua configuração de conjunto, com o *após* com pequeno déficit e, principalmente, pelo déficit significativo de *depois*.

3.6 PREPOSIÇÕES + DEMONSTRATIVOS

Resolvemos reunir as preposições aos demonstrativos a fim de ampliar o espectro de funcionais para a análise:

Figura 13. Contomad - Preposições + demonstrativos.

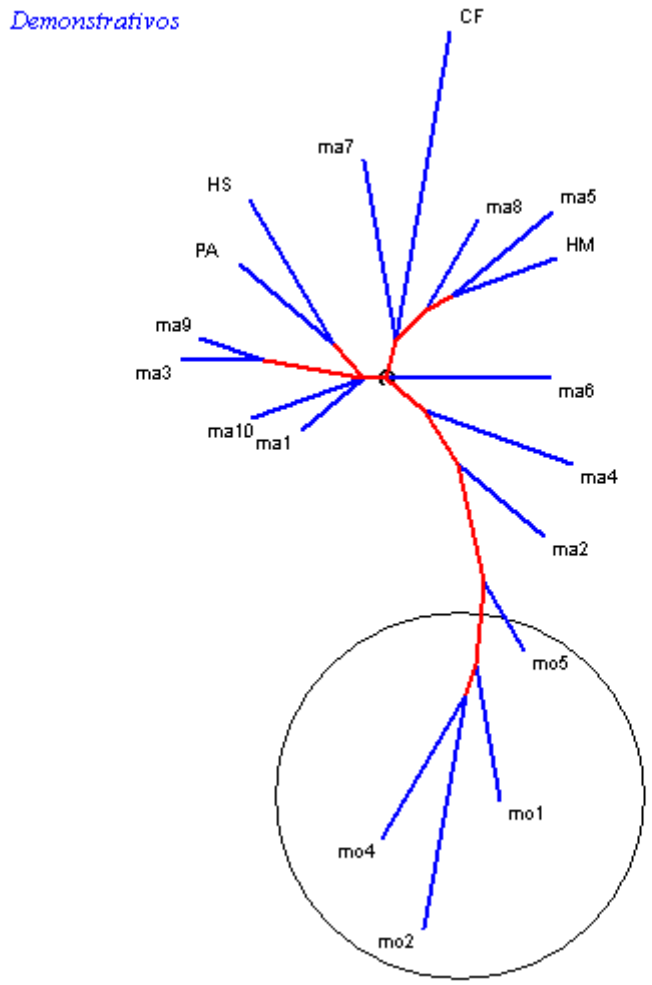


Esta é uma das análises em árvores que divide de forma mais nítida os grupos analisados. Os ramos e as folhas distribuem-se em três direções distintas. Na parte mais curva, à esquerda, temos os avulsos agrupados de forma peculiar: ma3 e ma9, que são quase gêmeos em tema e estrutura, partem de um ramo que divide exatamente ao meio os demais, quatro para cada lado.

Indo para a direita, em sentido horário, aparecem agrupados os volumes de contos, dois a dois por cronologia. E, o que mais nos interessa: no alto, *Madalena*, agrupada a Moreira, vai consolidando esta posição.

3.7 DEMONSTRATIVOS

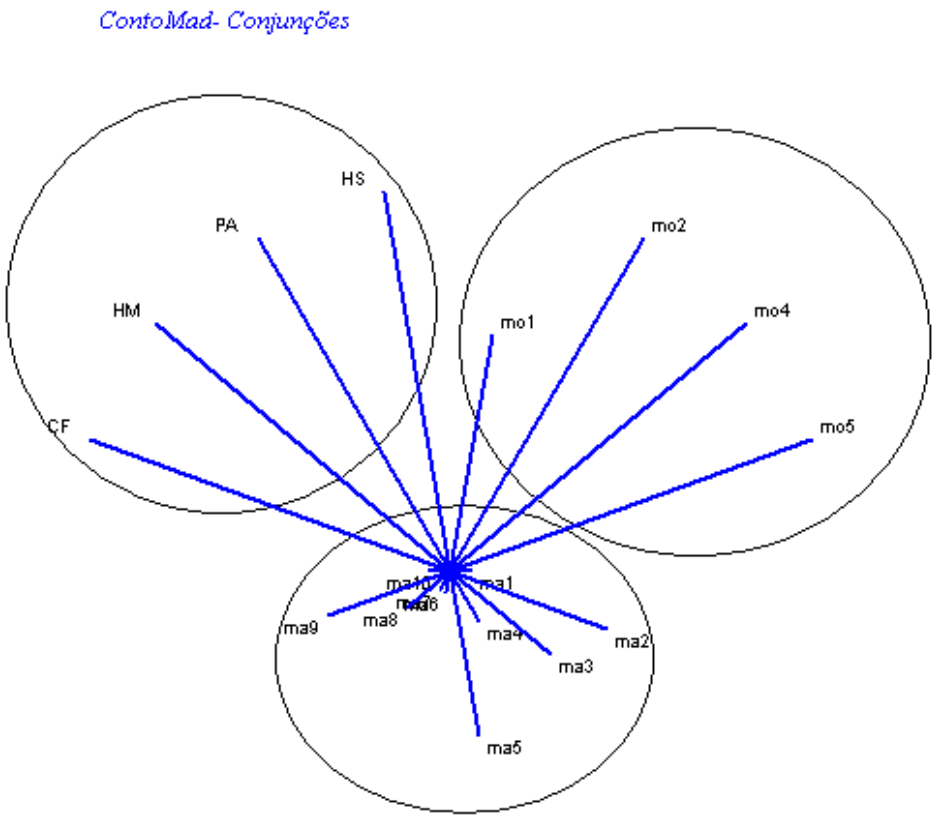
Figura 14. Contomad - Demonstrativos – estrutura.



Ao desconstruirmos um pouco mais nossa árvore, deixando apenas os demonstrativos, parece que alcançamos um dos elos perdidos entre *Madalena* e Moreira, que agora não deixa mais dúvidas sobre sua vizinhança ou grupo.

3.8 CONJUNÇÕES

Figura 15. Contomad – conjunções-estrutura (216 combinações conjunções).



Finalmente, chegamos ao último item de estrutura: as conjunções. A partir de 216 configurações diferentes em que a conjunção é a constante, conseguimos a mais distinta em termos de divisão. Esta árvore impressiona pela equidistância entre os três grupos. Os avulsos encontram-se na parte de baixo, com as suas diferenças marcadas, no caso, pelo comprimento das folhas. No sentido anti-horário, podemos ler os códigos dos volumes de contos de Machado exatamente na sequência cronológica. E, à direita, temos *Madalena* integrada ao grupo de Moreira. Os quatro, como o grupo de Machado, exatamente na sequência, seguindo a linha do tempo.

3.9 DADOS LEXICAIS

Este trabalho é focado na estrutura, no entanto, consideramos necessário, mesmo que de forma complementar, apresentar alguma amostra de dados lexicais – ainda que consideremos

a ressalva de que os dados lexicais são sensíveis ao tema, entre outros fatores. Para tal, optamos por um dos recursos mais importantes do *Hyperbase*: a análise da distância lexical.

Duas são as maneiras básicas de medir a distância entre dois textos: uma é calculada sobre as ocorrências e outra sobre o vocabulário. Optamos pelo vocabulário, que é mais individualizante, a partir de duas bases: Momamag e Contomad, nesta mesma ordem:

Figura 16. Base Momamag, distance lexical (V).

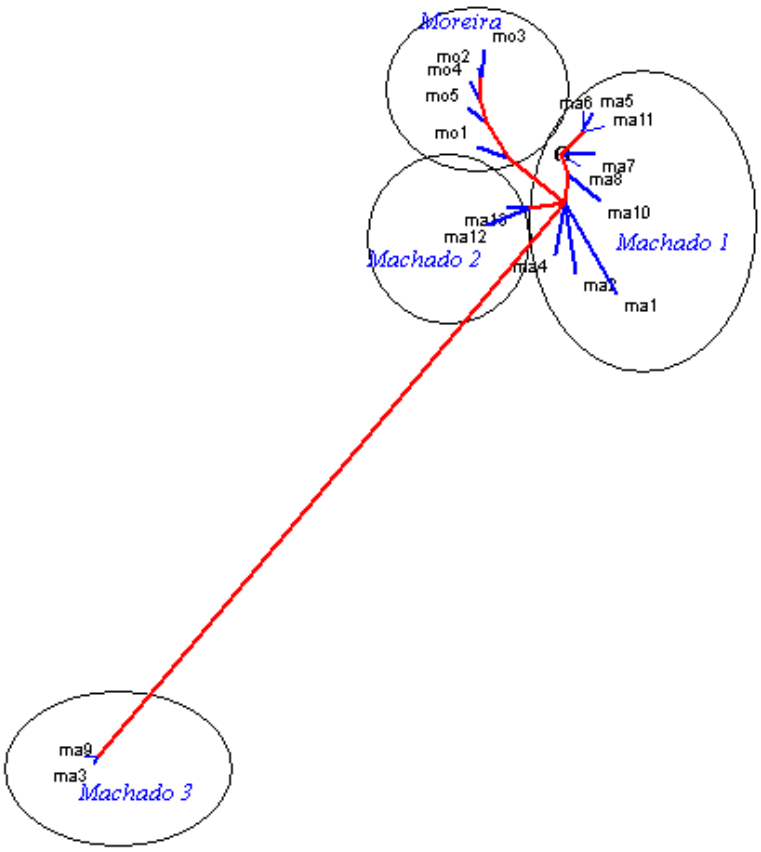
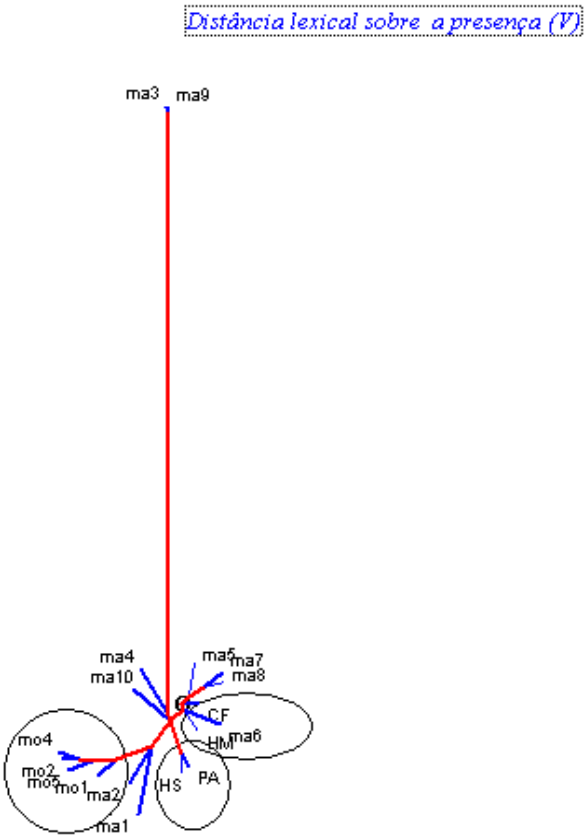


Figura 17. Base Contomad, distance lexical (V).



As duas árvores de distância lexical corroboram a hipótese de que autoria de *Madalena* pertence a Moreira de Azevedo.

O caso *Madalena* nunca recebeu um estudo sério do ponto de vista linguístico e estilístico a fim de autenticá-lo como de Machado ou Moreira. Obviamente, um conhecedor de Machado desconfia de um conto em que tudo termina bem, mas isto não basta, serve apenas de ponto de partida. E nada melhor para dar segurança a um trabalho que se pretende científico do que a linguagem dos números. Embora saibamos que eles não bastam, pois são vulneráveis a graus de manipulação, como acontece com outros métodos, e também a erros técnicos, a falhas de interpretação etc.

Ao final deste trabalho, não só constatamos que, de fato, o caminho da atribuição de autoria é tortuoso, como consideramos que é a combinação de técnicas e cálculos a melhor direção a tomar. Não acreditamos que haja um caminho universal, uma vez que cada obra, a despeito das constantes do gênero literário e da época a que pertence, é, antes de tudo, singular.

Quanto aos objetivos deste trabalho, só podemos afirmar que, ao menos estatisticamente, pelo resultados dos 37 gráficos e árvores apresentados, consideramos doze deles com resultado indefinido a respeito da autoria de *Madalena*. Ou seja, ao menos nossos números atribuem a Moreira de Azevedo a paternidade de *Madalena*.

REFERÊNCIAS

BRUNET, Etienne. Peut-on mesurer la distance entre deux textes? *Corpus*, Nice, n. 2, La distance intertextuelle, décembre, 2003. Disponível em: <https://*corpus*.revues.org/34>. Acesso em: 20 fev. 2017.

DUBROCARD, M.; LUONG, X. Problèmes d'attribution: application de quelques tests statistiques à different historiens latins, analyse arborée. *Proceedings of Vestal '99* - Venezia, San servolo, V.I.U. p. 293-300, 1999.

GARCÍA, A. M.; MARTÍN, J. C. Function words in authorship attribution studies. *Literary and Linguistic Computing***,** Oxford, v. 22, n. 1, p. 49-66, 2007.

GRIEVE, J. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing***,** Oxford, v. 22, n. 3, p. 49-66, 2007.

GUIRAUD, Pierre. *Essais de stylistique*. França: Klincksieck, 2000. p. 62.

HOCKEY, Susan. *Electronic texts in the humanities*. London, New York: Oxford University Press, 2004.

INGARDEN, Roman. *A Obra de Arte Literária*. Tradução de Albin E. Beau Maria da Conceição Puga e João F. Barrento. 2 ed. Lisboa: Fundação Calouste Gulbenkian, 1965.

JAUBERT, Anna. Des genres comme précurseurs de style. *Loxias 8*, Emergence et hybridation des genres, Nice, 1999. Disponível em: <http://revel.unice.fr/loxias/?id=99>. Acesso em: 20 fev. 2017.

JUOLA, P. *Authorship Attribution*. Boston: Now Publishers, 2006.

LABBÉ, Cyril; LABBÉ, Dominique. La distance intertextuelle. *Corpus*, n. 2, dez. 2003.

LEBART, L. Validité des visualisation de donnés textuelles. JADT 2004: *7es Journées internationales d'Analyse statistique de Données Textuelle*s, p. 709-715.

LUONG, Xuan; MACIEL, Carlos. Fréquences e repartition des mots dans um *corpus* de littérature brésilienne, in *Jadt 2002, 6es. Journées internationales d'Analyse statistique des Données Textuelles*.

LUONG, Xuan; MELLET, Sylvie. Mesures de distance grammaticale entre les textes. *Corpus*, Nice, n. 2, La distance intertextuelle, dez. 2003. Disponível em: <https://*corpus*.revues.org/34>. Acesso em: 20 fev. 2017.

LUONG, Xuan; NOVI, Michel. Répresentations arborées de données textuelles. In: *Méthodes quantitatives et informatiques dans l'étude des textes*: colloque international. Université de Nice, 5-8 juin 1985, en hommage à Charles Muller / Slatkine - Champion, p.577- 586, 1986.

MACIEL, C. A. A. *Richesse et evolution du vocabulaire d'Érico Veríssimo* (1905-1975 – Porto Alegre, Brésil). Paris-Genéve: Champion-Slaktine, 1986. p. 55.

MASSA, J. M. *Dispersos de Machado de Assis*. Rio de Janeiro: INL, 1965.

McMENAMIN, Gerald R. *Forensic linguistics: advances in forensic stylistics*. Boca Raton: CRC Press LLC, 2000.

SOUSA, J. Galante de. *Bibliografia de Machado de Assis.* Rio de Janeiro: INL, 1955.

Recebido: 11 de julho de 2017
Aceito: 15 de julho de 2017