doi: http://dx.doi.org/10.5007/1984-8412.2022.e79086

INDIVIDUAÇÃO DE AUTORIA E IDENTIFICAÇÃO DE ESTILO: ANÁLISE DE OBRAS LITERÁRIAS COM AUXÍLIO DO R

INDIVIDUACIÓN DE AUTORÍA E IDENTIFICACIÓN DE ESTILO: ANÁLISIS DE OBRAS LITERÁRIAS CON R

INDIVIDUATION OF AUTHORSHIP AND STYLE IDENTIFICATION: ANALYSIS OF LITERARY WORKS CARRIED WITH R

Luis Filipe Lima e Silva*
Larissa S. Ciríaco**
Universidade Federal de Minas Gerais

RESUMO: Este artigo soma-se aos trabalhos disponíveis sobre Processamento de Língua Natural ao fornecer uma demonstração de como linguagens de programação como o R (R CORE TEAM, 2020) podem ser úteis na detecção de autoria e na identificação do estilo do autor em obras literárias. Foram selecionados dois autores e duas obras de cada, a saber: *The Adventures of Tom Sawyer* (1876) e *Adventures of Huckleberry Finn* (1884), do autor Mark Twain (1835-1910), e *Typee: A Peep at Polynesian Life* (1846) e *Omoo: A Narrative of Adventures in the South Seas* (1847), do autor Herman Melville (1819-1891). Posteriormente, os dados foram analisados seguindo a mesma metodologia de Eder *et al.* (2016), a fim de testar a eficácia do pacote *stylo* e aplicar os métodos de Análise de Componentes Principais, Análise de *Cluster* e Árvore de Consenso. Os resultados apontaram que cada um dos métodos testados conseguiu distinguir as obras dos autores, evidenciando-se, assim, a eficácia do pacote utilizado. Além disso, realiza-se uma

* Possui doutorado (2020) e mestrado (2016) em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG). E-mail: luisf.1397@gmail.com.

** Professora de Linguística da Universidade Federal de Minas Gerais (UFMG). E-mail: larissaciriaco@ufmg.br.

análise estilométrica baseada nos métodos de Zeta de Craig e *Rolling Delta*. Para este último, utilizaram-se obras de dois autores de língua alemã, Frank Kafka e Heinrich von Kleist. Os resultados apontaram uma semelhança estilística de von Kleist, sobretudo, na primeira obra de Kafka. Adicionalmente, o método *Rolling Delta* foi usado para explorar uma análise feita por Juola (2013ª, 2013b) a respeito de uma obra de J. K. Rowling escrita sob o pseudônimo de Robert Galbraith.
PALAVRAS-CHAVE: Detecção de autoria. Análise estilométrica. R.

RESUMEN: Este artículo se suma a los trabajos disponibles sobre procesamiento del lenguaje natural al proporcionar una demostración de cómo los lenguajes de programación como R (R CORE TEAM, 2020) pueden ser útiles para detectar la autoría e identificar el estilo del autor en obras literarias. Se seleccionaron dos autores y dos obras de cada uno, a saber: *The Adventures of Tom Sawyer* (1876) y *Adventures of Huckleberry Finn* (1884) del autor Mark Twain (1835-1910), y *Typee: A Peep at Polynesian Life* (1846) y *Omoo: A Narrative of Adventures in the South Seas* (1847) del autor Herman Melville (1819-1891). Posteriormente, los datos se analizaron utilizando la misma metodología que Eder et al. (2016), con el fin de probar la efectividad del paquete *stylo* y aplicar los métodos de Análisis de Componentes Principales, Análisis de *Cluster* y Árbol de Consenso. Los resultados mostraron que cada uno de los métodos probados fue capaz de distinguir los trabajos de los autores, evidenciando así la efectividad del paquete utilizado. Además, se realiza un análisis estilométrico basado en los métodos de Zeta de Craig y *Rolling Delta*. Para esto último, se utilizaron obras de dos autores de habla alemana, Frank Kafka y Heinrich von Kleist. Los resultados apuntan a una similitud estilística de von Kleist, sobre todo, en la primera obra de Kafka. Además, el método *Rolling Delta* fue utilizado para explorar un análisis de Juola (2013ª, 2013b) sobre una obra de J. K. Rowling escrita bajo el seudónimo de Robert Galbraith.
PALABRAS-CLAVE: Detección de autoría. Análsis estilométrico. R.

ABSTRACT: This paper adds to the works available on Natural Language Processing by providing a demonstration of how programming languages such as R (R CORE TEAM, 2020) can be useful in detecting authorship and identifying the style of the author in literary works. Two authors and two works each were selected, namely: The Adventures of Tom Sawyer (1876) and Adventures of Huckleberry Finn (1884) by author Mark Twain (1835-1910), and Typee: A Peep at Polynesian Life (1846) and Omoo: A Narrative of Adventures in the South Seas (1847) by author Herman Melville (1819-1891). Subsequently, the data were analyzed following the same methodology as Eder et al. (2016), in order to test the effectiveness of the *stylo* package and apply the Principal Component Analysis, Cluster Analysis and Consensus Tree methods. The results showed that each of the tested methods was able to distinguish the works of the authors, thus evidencing the effectiveness of the package used. In addition, a stylometric analysis is performed based on Craig's Zeta and Rolling Delta methods. For the latter, works by two German-speaking authors, Frank Kafka and Heinrich von Kleist, were used. The results pointed to a stylistic similarity of von Kleist, especially in Kafka's first work. Additionally, Rolling Delta was used to explore an analysis carried by Juola (2013a, 2013b) regarding a work by J. K. Rowling written under the pseudonym of Robert Galbraith.
KEYWORDS: Authorship detection. Stylometric analysis. R.

1 INTRODUÇÃO

O objetivo deste artigo é contribuir para os estudos e os trabalhos (cf. SILGE; ROBINSON, 2017; JOSHI, 2019; BOEHMKE; GREENWELL, 2019, entre outros) que mostram como o R (R CORE TEAM, 2020), um ambiente de programação, pode contribuir para pesquisas na área de Processamento de Língua Natural (PLN), especialmente em casos em que se precisa verificar a autoria de textos[1] – tema comumente relacionado também à área de Linguística Forense[2]. Para isso, foram utilizados o pacote *stylo* (EDER *et al*. 2016) e os seguintes métodos: Análise de Componentes Principais, Análise de *Cluster*, Árvore de Consenso, Zeta de Craig e *Rolling Delta*. O artigo traz um passo a passo de como utilizar cada método e explica as vantagens e as desvantagens de cada um,

[1] Vale mencionar que o R não é o único recurso disponível para esse fim. Há outras plataformas para identificação de autoria, como a linguagem de programação *Python* e pacotes como *Scikit-Learn*.

[2] É importante deixar claro que os algoritmos utilizados não se aplicam diretamente à Linguística Forense, cujos recursos de identificação de autoria atualmente são baseados em redes neurais e superam, em precisão, os algoritmos apresentados neste artigo. Ainda assim, para a Linguística Forense, há também o pacote SimilaR, planejado para detectar plágio em código de programas em R (cf. BARTOSZUK; GAGOLEWSKI, 2020).

replicando o estudo de Eder *et al.* (2016) com outras obras[3]. Através de um banco de dados de obras literárias, será mostrado o que se conhece por análise de autoria, um campo de estudos relacionado ao PLN por usar dados da língua natural para gerar suposições sobre seus usuários. Os recursos computacionais empregados nesta área permitem a identificação de autor anônimo, detecção de plágio e de *ghost writer*, identificação de estilo de escrita etc. (QIAN *et al.* 2017; STAMATATOS, 2009). Várias aplicações práticas emergem desse tipo de análise. Ela tem sido aplicada a trabalhos literários, de inteligência artificial, direito civil e criminal, bem como à área de computação forense (ZHANG *et al.* 2014). Na seção seguinte, será feita uma breve apresentação do R. Na seção 3, será demonstrada a aplicação do R para análises de aferição de autoria em duas frentes: em primeiro lugar, as análises de PCA, *Cluster*, e Árvore de Consenso[4] mostram a individuação de autoria; em segundo lugar, são mostradas as análises de Zeta de Craig e *Rolling Delta*, que servem mais especificamente a um propósito de identificação de estilo, tanto do ponto de vista do léxico utilizado (Zeta de Craig) quanto do ponto de vista da influência de outras obras literárias na escrita do autor (*Rolling Delta*). Adicionalmente, será mostrada uma análise com o método de *Rolling Delta* replicando o estudo de Juola (2013a; 2013b) a respeito de uma obra de J. K. Rowling escrita sob o pseudônimo de Robert Galbraith.

2 APRESENTANDO O R

O R é uma linguagem de programação livre e um local de desenvolvimento de computação, cálculos estatísticos e gráficos para interpretação de dados. Como programa, o R pode ser estendido facilmente por meio de pacotes disponíveis na internet através da rede de *sites* CRAN (*The Comprehensive R Archive Network*, 2020) ou da coleção Tidyverse (2020), ambos disponíveis *online*. Já o RStudio é um ambiente de desenvolvimento integrado do R, ou seja, um *software* profissional à disposição dos pesquisadores que trabalham com ciência de dados e de outras disciplinas cujos trabalhos envolvam tratamento e análise estatística de dados. Por meio do RStudio, é possível não só desenvolver o trabalho, mas também compartilhá-lo em ampla escala. O R fornece, ainda, uma variedade de técnicas estatísticas e gráficas, como modelagem linear e não linear, testes estatísticos clássicos, análise por meio de *clustering*, *time-series* ou modelagem de séries temporais etc. O R também tem seu próprio formato LaTeX, que pode ser usado para produzir documentos de vários modelos.

Uma das vantagens em se utilizar o R está na facilidade com que se pode criar gráficos bem desenhados e de alta qualidade, inclusive para o padrão de publicação. Um pacote muito utilizado é o ggplot2, que, além de construir gráficos de boa qualidade, é fácil de customizar. Sendo assim, o R é uma excelente ferramenta de visualização de dados. Segundo o *site* do R (2020), o programa inclui: um recurso para estoque e manuseio efetivo dos dados, uma coleção ampla e coerente de ferramentas para análise de dados, facilidades gráficas para análise de dados (que podem ser apresentadas na tela ou em cópia física) e uma linguagem de programação simples, efetiva e bem desenvolvida que inclui inúmeras funções. Outra vantagem é a facilidade com que se pode descobrir algo sobre o *software* ou tirar uma dúvida sobre um problema de programação, seja pelo *Google*, seja através do *site* de perguntas e respostas sobre programação *StackOverflow* (2020). Nesse *site*, as questões costumam ser respondidas rapidamente por outros usuários do R[5].

O R pode ser utilizado para análise de dados em diversas disciplinas científicas, como a física, a química, a biologia, as ciências sociais e a linguística. Na linguística, o R tem auxiliado os pesquisadores a interpretar dados extraídos de *corpora* para análises linguísticas, dados resultantes de experimentos comportamentais e cronométricos para análises psicolinguísticas e linguístico-cognitivas, entre outros. Neste artigo, serão focalizadas análises a partir do R para interpretar dados de obras literárias com a finalidade de identificação de autoria e de estilo.

3 APLICAÇÃO DO R NA AFERIÇÃO DE AUTORIA E IDENTIFICAÇÃO DE ESTILO

[3] O objetivo de replicar o estudo com outras obras é testar a eficácia do pacote, isto é, até que ponto podemos ter resultados eficazes utilizando os mesmos métodos.

[4] Trabalhos em ciência de dados também utilizam o termo "árvore de decisão". Neste artigo, optamos por "árvore de consenso".

[5] O realce das vantagens do R tem um objetivo meramente descritivo, tendo em vista que ele foi o software escolhido para o trabalho deste artigo. Obviamente, outros softwares podem exceder o R em muitos aspectos. Por exemplo, Java e C++ são opções com maior performance; Python também é muito utilizado em processamento da linguagem natural, e especialmente para *deep learning*; etc.

Nesta seção, será apresentada uma demonstração do uso do R na pesquisa com dados de língua natural. Nosso objetivo não é discutir os parâmetros estilométricos para a identificação de autoria ou problematizar os métodos estatístico-computacionais conduzidos para o processo de identificação. O objetivo desta seção é apenas mostrar como o uso do R pode ser útil para a pesquisa com esse tipo de dados e, mais especificamente, nas análises de autoria e de estilo.

Inicialmente, exemplifica-se a possibilidade de *individuação de autoria*, isto é, como é possível a aplicação de um método que mostre que os textos considerados para análise sejam obras distintas de um mesmo autor (3.1, 3.2, 3.3). Posteriormente, demonstra-se como é possível identificar o estilo do autor (3.4, 3.5). O algoritmo será alimentado com quatro obras de dois autores diferentes, a fim de que seja possível distingui-las. As obras selecionadas foram *The Adventures of Tom Sawyer* (1876) e *Adventures of Huckleberry Finn* (1884), do autor Mark Twain (1835-1910), e *Typee: A Peep at Polynesian Life* (1846) e *Omoo: A Narrative of Adventures in the South Seas* (1847), do autor Herman Melville (1819-1891). Essas obras foram extraídas do sítio do *Project Gutenberg* (https://www.gutenberg.org/) – uma biblioteca digital que conta com mais de 38.000 livros em diferentes línguas, tanto em formato de texto quanto em formato de livros-áudio. Contar com a obra digitalizada em formato txt é essencial, pois o pacote utilizado para o estudo lê apenas textos nesse formato. O pacote selecionado para a análise foi o *stylo* (EDER *et al.*, 2016), que permite análises estatísticas exploratórias do estilo de escrita em línguas como inglês, alemão, francês, latim, espanhol, holandês, polonês e húngaro, além de ser compatível com outros sistemas de escrita, como coreano, chinês, japonês, árabe, hebraico, cóptico e grego.

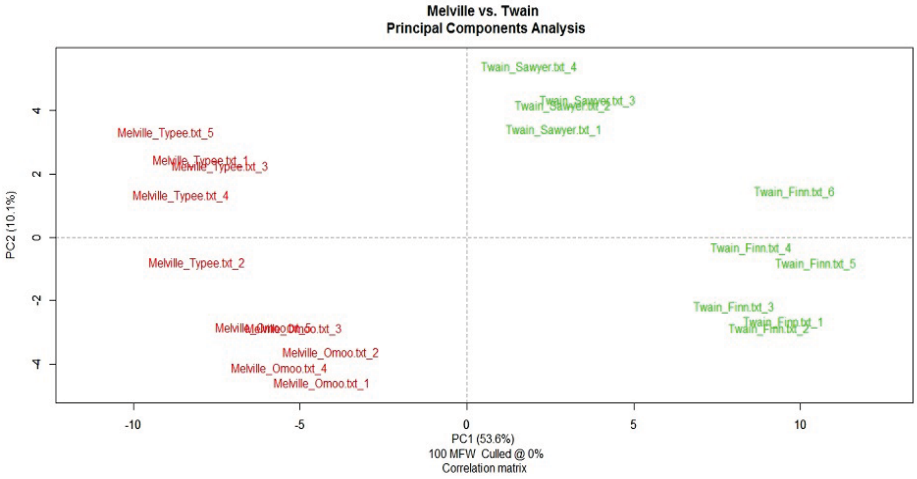
3.1 PCA (*PRINCIPAL COMPONENT ANALYSIS*)

O primeiro método que será mostrado é a Análise de Componentes Principais (*Principal Components Analysis* – PCA). O PCA é um procedimento de análise multivariado que permite, entre outras funções, revelar a existência de *relações entre amostras* (cf. JOLLIFFE, 2002). A análise agrupa as amostras de acordo com suas variâncias, isto é, as amostras são agrupadas de acordo com seu comportamento dentro de uma população, que é representado pela variação de um conjunto de características que as determina. Essa variação das características da amostra permite que o PCA agrupe os dados e forneça a distribuição das amostras num gráfico bidimensional. O PCA extrai informação de uma tabela de dados multivariados e expressa a informação como um conjunto de novas variáveis denominadas componentes principais. Tais componentes principais correspondem a uma combinação linear das variáveis originais. Em última análise, o PCA reduz a dimensionalidade de dados multivariados em dois ou três componentes principais.

O PCA é um método útil quando as variáveis são correlacionadas, pois a correlação indica redundância e, a partir dessa redundância, o PCA foi usado para reduzir as variáveis originais no menor número possível de novas variáveis ou componentes principais, de modo a explicar a maior parte da variância nas variáveis originais. De forma simplificada, o PCA identifica um padrão escondido num conjunto de dados, identifica variáveis correlacionadas e reduz a dimensionalidade dos dados removendo a redundância.

O pré-processamento dos dados envolveu a alimentação das obras em formato txt UTF-8 Unicode no *software* R, a tokenização das obras, isto é, a divisão do texto em unidades contáveis, tais como *tokens* de palavras. Os pronomes pessoais foram excluídos, uma vez que eles tendem a estar correlacionados com um assunto específico ou gênero de um texto (cf. PENNEBAKER, 2011). Além disso, as palavras foram convertidas em trigramas, uma vez que se deseja extrair unidades mensuráveis e de alta frequência para a identificação de autoria (cf. EDER, 2011). Posteriormente, as obras foram fatiadas em amostras como parte de um processo de avaliação da coerência estilística, sendo que tais fatias comportaram amostras não sobrepostas de 20.000 palavras cada. Adicionalmente, foi criada uma tabela de frequência dos 3000 traços mais frequentes do *corpus*. Seguiu-se com a extração de um vetor para cada amostra contendo as frequências relativas dos traços mais frequentes combinadas numa tabela, que foi usada para análise estatística. Por meio de um critério de seleção, especificou-se uma porcentagem de ocorrência das palavras nas amostras como ponto de corte para análise, ou seja, palavras que não ocorriam, pelo menos, com uma taxa de 80% nas amostras foram ignoradas. O gráfico 1 mostra o resultado da análise do PCA utilizando as cem palavras mais frequentes de cada obra. Os dois primeiros componentes PC1 e PC2 explicam 63,7% da variância dos dados. Se observarmos o lado esquerdo da figura 1, veremos que os três primeiros componentes explicam cerca de 70% da variância dos dados, que é uma porcentagem de grande aceitabilidade. O lado direito da figura 1 é apenas outra forma de ilustrar a análise do PCA. No gráfico 1, cada barra corresponde a um componente principal.

Gráfico 1: PCA Melville vs. Twain



Fonte: Elaboração própria

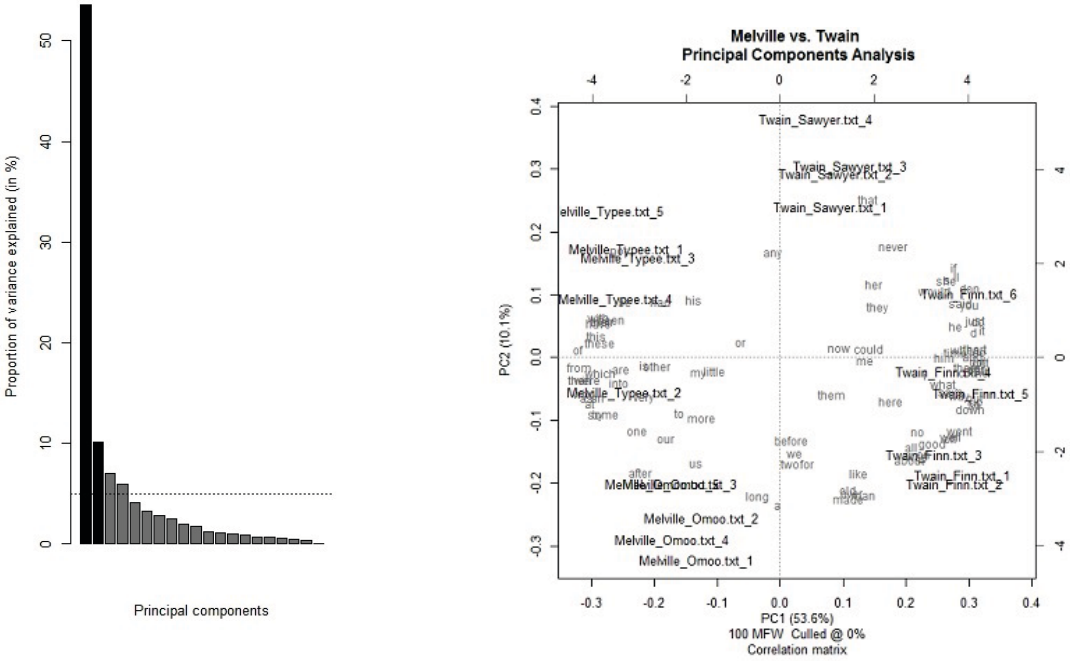


Figura 1 : PCA Melville vs. Twain

Fonte: Elaboração própria

Observe que, no gráfico 1, as obras são representadas por números que indicam uma fatia de amostra de cada obra. Essas amostras foram agrupadas segundo suas características de correlação, sendo que as que foram mais correlacionadas foram agrupadas mais próximas umas das outras. Esse gráfico deve ser interpretado da seguinte maneira: correlações positivas são agrupadas juntas, ao passo que correlações negativas são distribuídas nos quadrantes opostos do gráfico. Quanto mais distantes do centro do gráfico mais bem representadas são tais variáveis. Nota-se que as amostras das obras estão bem espaçadas, distantes do ponto de origem do gráfico. Portanto, é possível afirmar que elas contribuem para os componentes principais. Além disso, as correlações encontradas entre as obras dos dois autores, que estão agrupadas em lados opostos do gráfico, permitem dizer que elas são obras de dois autores distintos. Isso é confirmado pelo fato de haver cores diferentes para os dois autores. Observe que as obras de Herman Melville estão representadas em vermelho, ao passo que as obras de Mark Twain estão representadas em verde. Observe, também, que cada obra está representada em um quadrante específico do gráfico, isto é, elas não se agruparam num mesmo espaço, o que permite dizer que

se trata de obras distintas de dois autores diferentes. Portanto, por meio da Análise de Componentes Principais, foi possível mostrar a identificação de autoria e de obras distintas de um mesmo autor.

3.2 *CLUSTER*

O segundo método que será mostrado é Análise de *Cluster* (cf. HENNIG *et al*., 2016). A Análise de *Cluster* é um método multivariado que consiste em agrupar objetos com base em um conjunto de variáveis *segundo suas semelhanças*, de forma que eles pertençam a um mesmo grupo ou *cluster*. Um *cluster* se diferencia de outros *clusters* devido ao agrupamento de outros conjuntos de variáveis semelhantes. Em outras palavras, os *clusters* classificam objetos semelhantes entre si, mas que se distinguem de outros como grupo. Há diferentes métodos de análise de *cluster*, como os métodos hierárquicos e os não hierárquicos. Os primeiros consistem em "[...] agregação sucessiva ou divisão das observações e de seus subconjuntos. Resultante desse tipo de procedimento, surge uma estrutura em forma de árvore, que é conhecida como dendrograma" (WIERZCHOŃ; KŁOPOTEK, 2018, p. 29)[6]. Já os métodos não hierárquicos atribuem cada variável a um *cluster* num espaço multidimensional sem mostrar as interrelações entre os objetos. No caso em análise, os objetos a serem agrupados em *clusters* hierárquicos são as amostras fatiadas das obras analisadas, assim como ocorreu com o PCA (cf. seção 3.1). As mesmas obras utilizadas para o PCA foram usadas para essa análise de *cluster*. Dessa vez, a análise se baseia nas 624 palavras mais frequentes de cada obra.

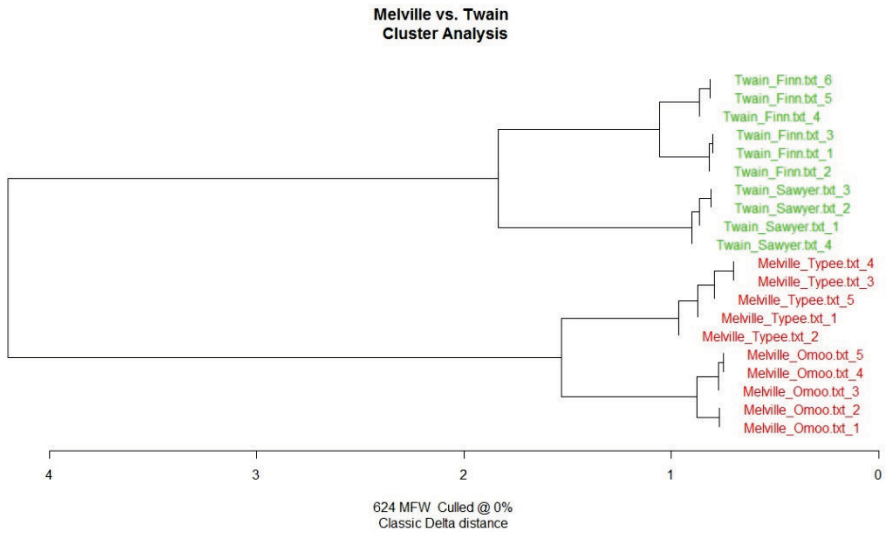
Para a implementação da análise de *cluster*, é preciso ter um método de medir a distância entre as observações, sendo que o tipo de medida dependerá do tipo de dado que se considera para análise. A medida de distância utilizada para a análise de *cluster* utilizada aqui é a Delta clássica (BURROWS, 2002). Essa medida, utilizada para a atribuição de autoria, é definida como "[...] a média das diferenças absolutas entre escores-z para um mesmo conjunto de variáveis de palavras em um determinado grupo de texto e os escores-z para o mesmo conjunto de variáveis de palavras em texto alvo" (BURROWS, 2002, p. 271)[7]. Essa medida pressupõe que existe um conjunto de textos de comparação cujos escores-z serão calculados. Os escores-z são calculados com base na média e no desvio padrão das frequências das palavras no *corpus* de comparação. A medida Delta é calculada "[...] entre o texto alvo e cada um de um conjunto de textos candidatos (geralmente compreendendo o corpus de comparação), e o alvo é atribuído ao autor do texto candidato com o escore Delta mais baixo" (ARGAMON, 2008, p. 131)[8]. Abaixo, no gráfico 2, é mostrado o dendrograma formado a partir da análise de *cluster* das quatro obras utilizadas neste estudo.

[6] Tradução nossa do original: "[…] successive aggregation or division of the observations and their subsets. Resulting from this kind of procedure there is a tree-like structure, which is referred to as dendrogram".

[7] Tradução nossa do original: "[…] the mean of the absolute differences between z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text".

[8] Tradução nossa do original: "[…] between the target text and each of a set of candidate texts (generally comprising the comparison corpus), and the target is attributed to the author of the candidate text with the lowest Delta score".

Gráfico 2: Dendrograma Melville vs. Twain



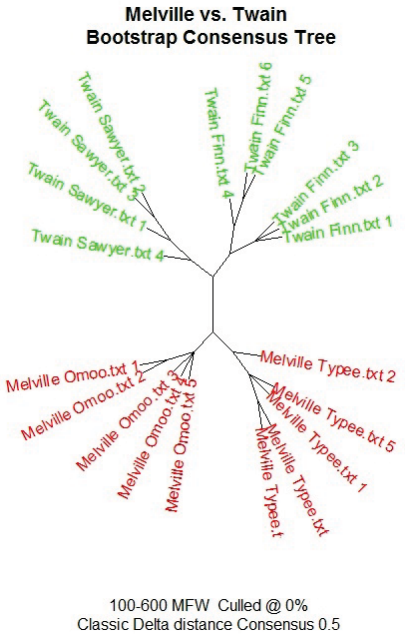
Fonte: Elaboração própria

É possível observar que há duas grandes bifurcações que separam os dois autores, e, dentro delas, outras duas bifurcações menores e assim sucessivamente. Observe que isso pode ser notado através das cores diferentes que representam as obras de Mark Twain, em verde, e as obras de Herman Melville, em vermelho. As duas bifurcações dentro das duas primeiras bifurcações separam as obras de cada autor. Isso é verificado notando que essas bifurcações separam inicialmente *Adventures of Huckleberry Finn* de *The Aventures of Tom Sawyer*, bem como, mais abaixo, *Typee: A Peep at Polynesian Life* de *Omoo: A Narrative of Adventures in the South Seas*. As bifurcações subsequentes que estão dentro dessas quatro bifurcações representam cada amostra fatiada das obras, assim como foi representado no PCA. Observa-se, portanto, que a aplicação de outro método de análise gerou os mesmos resultados do método anterior.

3.3 ÁRVORE DE CONSENSO

O terceiro método é a Árvore de Consenso (*Consensus tree*). O pressuposto desse método é que os textos podem ser representados como nódulos de uma rede, e as relações que possam ser estabelecidas dentro dessa rede são ligadas por meio dos nódulos. Esse procedimento tem como objetivo a identificação de alvos vizinhos ao mesmo tempo que *extrai os padrões* mais robustos que definem os sinais de autoria, filtrando as semelhanças textuais que são identificadas como menos relevantes (EDER, 2017). O método funciona do seguinte modo: ele assume que agrupamentos tendem a reaparecer a partir de um grande número de *snapshots*, por exemplo, para as 100, 200, 300 etc. palavras mais frequentes, sendo que as potenciais semelhanças que possam emergir são consideradas acidentais. O objetivo do método é, então, apreender os padrões robustos dos dados num conjunto de *snapshots* que são gerados. Posteriormente, o método produz certa quantidade de dendrogramas virtuais que servirão de base para uma avaliação da robustez dos agrupamentos gerados por tais dendrogramas. A frequência de ligações que um agrupamento gera entre amostras de obras ou entre diferentes obras será a fonte para a produção de um gráfico de consenso. Dito de outro modo, vários dendrogramas virtuais 'votam' para aferir as ligações mais robustas de modo que tal procedimento resume a informação processada no agrupamento das semelhanças de partes específicas das obras, gerando o gráfico de consenso. As ligações mostradas no gráfico 3, de consenso, não dizem respeito a distâncias estilométricas entre as amostras das obras. Elas indicam tão somente o grau de concordância ou a força do consenso, ou, ainda, a repetição de uma série de dendrogramas virtuais. Para a presente análise das obras de Mark Twain e Herman Melville, os *snapshots* foram computados tomando, inicialmente, as 100 e alcançando até as 600 palavras mais frequentes, usando como medida de distância a Delta clássica, a mesma medida que foi realizada para a análise de *cluster*.

Gráfico 3: Árvore de consenso Melville vs. Twain



Fonte: Elaboração própria

Pode-se observar que a árvore de consenso conseguiu identificar não só os dois autores distintos, representados por cores particulares, mas também as quatros obras diferentes. Na parte superior, há dois grandes galhos compostos por alguns nódulos separando amostras diferentes de cada obra, sendo que esses galhos identificaram as obras do autor Mark Twain. Na parte inferior, há também dois grandes galhos juntamente com alguns nódulos separando amostras diferentes das obras do autor Herman Melville. A aplicação desse método também gerou os mesmos resultados na aferição de autoria, assim como nos dois outros métodos apresentados previamente. É importante notar, também, que *não só os autores foram identificados, mas também cada obra foi individualizada em amostras coerentes com os dados alimentados para a análise*. Pode-se concluir que a aplicação do PCA, da análise de *cluster* e da árvore de consenso foi satisfatória para o objetivo proposto, porque estas conseguiram singularizar com um alto nível de acerto as quatro obras dos dois autores tomados como exemplo para a aplicação dos métodos.

A seguir, exemplifica-se a possibilidade de *identificação de estilo de autoria*.

3.4 ZETA DE CRAIG

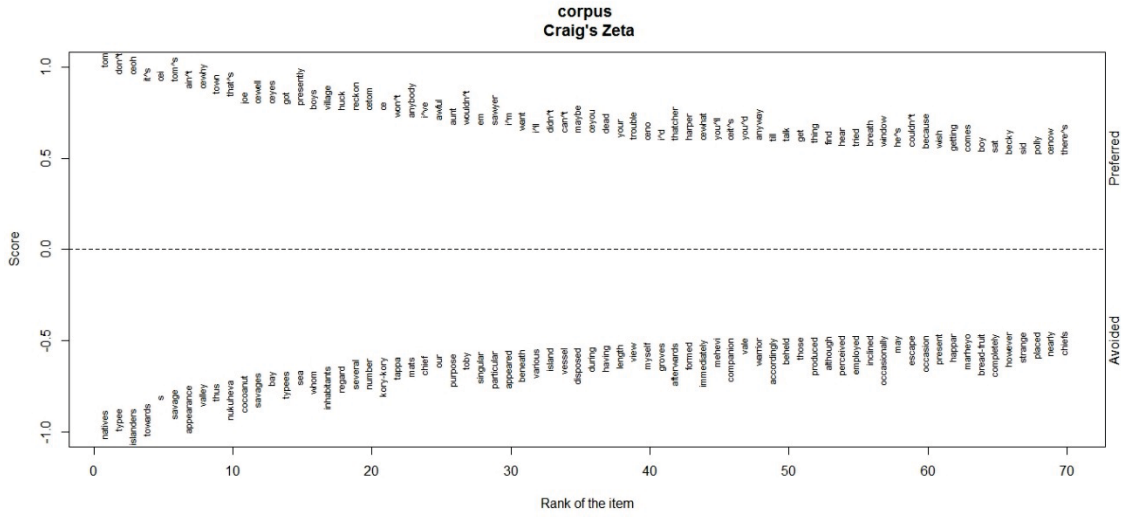
O pacote *stylo* também oferece um tipo de análise que mostra as *palavras mais usadas pelos autores* de cada obra. Esse procedimento é conhecido como Zeta de Craig (CRAIG; KINNEY, 2009). A análise Zeta foi desenvolvida por Burrows (cf. BURROWS, 2005; 2006; HOOVER, 2007a, 2007b, 2008). A versão utilizada neste estudo é, contudo, a que foi desenvolvida alternativamente por Craig. É importante mencionar que não se trata de uma análise das palavras-chave como é comumente feito em estudos de linguística de *corpus*. O Zeta de Craig compara as obras de dois autores diferentes e computa as palavras que um autor usa consistentemente, mas que o outro autor usa menos frequentemente ou evita usar, criando uma lista com um conjunto das palavras mais marcadas e outro com as palavras 'anti-marcadas'[9]. O Zeta de Craig exclui as palavras extremamente frequentes que, geralmente, são usadas numa análise comum de frequência, e concentra-se no meio do espectro de frequência das palavras. O procedimento se inicia com as duas obras divididas em seções de um mesmo tamanho. Então, passa-se a comparar quantas seções de cada autor contêm as respectivas palavras, de forma a ignorar as frequências das palavras, mas concentrando-se na consistência de sua aparição. De acordo com Hoover (2010, p. 1), o ponto central do método "[...] é que ele combina a proporção das seções de um autor em que cada palavra

[9] Embora Hoover (2010) não explique o significado do termo 'anti-marcado', acreditamos que ele se refira justamente às palavras evitadas por certo autor.

ocorre com a proporção das seções do outro autor do qual está ausente em uma única medida de distinção para cada palavra"[10]. Com essa classificação, duas listas de palavras são geradas, sendo uma favorecida pelo primeiro autor e evitada pelo segundo, e a outra favorecida pelo segundo autor e evitada pelo primeiro. Eder *et al.* (2016) interpretam o resultado gerado pela aplicação do Zeta de Craig como os traços de vocabulário comum entre um e outro autor, analisando o resultado de forma a inferir características do universo lexical a partir das listas geradas para as primeiras vinte palavras.

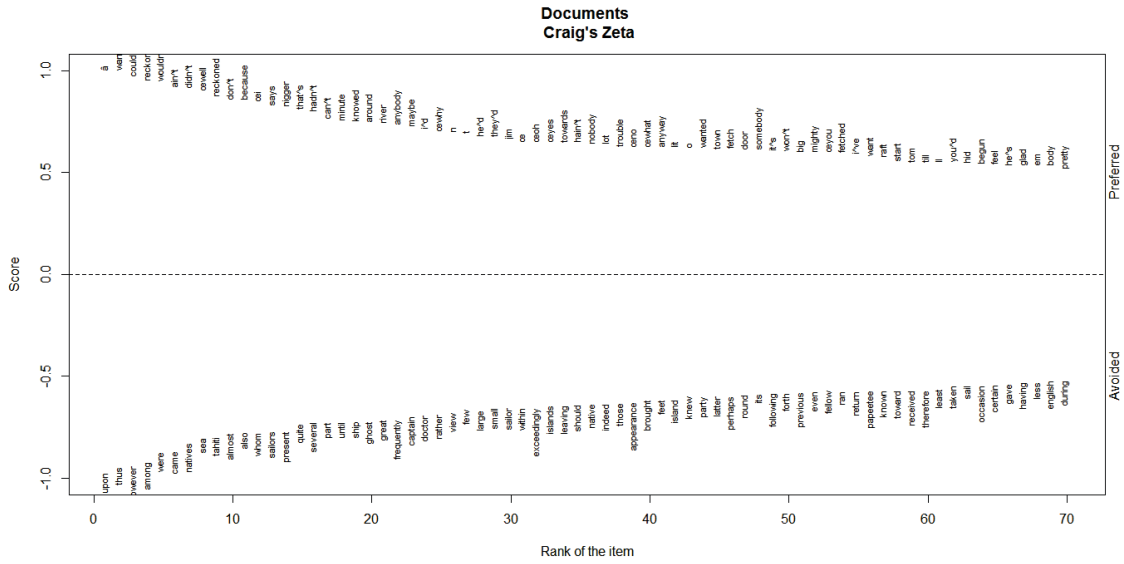
Para este estudo, foram computadas duas vezes o Zeta de Craig, uma comparando *The Adventures of Tom Sawyer*, de Mark Twain, com *Typee: A Peep at Polynesian Life*, de Herman Melville, e *Adventures of Huckleberry Finn* com *Omoo: A Narrative of Adventures in the South Seas*, dos mesmos respectivos autores. A aplicação desse procedimento gerou dois gráficos que podem ser visualizados abaixo. Apresentam-se, em seguida, no quadro 1, as vinte primeiras palavras da análise.

Gráfico 4: Zeta de Craig: Tom Sawyer vs. Typee



Fonte: Elaboração própria

Gráfico 5: Zeta de Craig: Huckleberry Finn vs. Omoo



Fonte: Elaboração própria

[10] Tradução nossa do original: "[…] is that it combines the ratio of the sections by one author in which each word occurs with the ratio of the sections by the other author from which it absent into a single measure of distinctiveness for each word".

Quadro 1: Vinte primeiras palavras características de cada obra

	Palavras preferidas	**Palavras evitadas**
The Adventures of Tom Sawyer vs. *Typee: A Peep at Polynesian Life*	Tom, don't, oh, it's, I, Tom's, ain't, why, town, that's, Joe, well, yes, got, presently, boys, village, huck, reckon, won't.	Natives, typee, islanders, towards, savage, appearance, valley, thus, nukuheva, cocoanut, savages, bay, typees, sea, whom, inhabitants, regard, several, number, kory-kory.
Adventures of Huckleberry Finn vs. *Omoo: A Narrative of Adventures in the South Seas*	A, warn't, couldn't, reckon, wouldn't, ain't, didn't, well, reckoned, don't, because, I, says, nigger, that's, hadn't, can't, minute, knowed, around.	Upon, thus, however, among, were, came, natives, sea, Tahiti, almost, also, whom, sailors, present, quite, several, part until, ship, ghost

Fonte: Elaboração propria

Pode-se notar que a análise mostra palavras gramaticais nas obras dos dois autores, contudo esse tipo de vocabulário aparece com predominância nas obras de Mark Twain. O vocabulário das obras de Herman Melville é caracterizado por palavras do ambiente marítimo, bem como referente aos aspectos do arquipélago onde a história se passa. Em última análise, o Zeta de Craig pode ser uma ferramenta útil para revelar o vocabulário preferido na obra de cada autor.

3.5 ROLLING DELTA - ANÁLISE ESTILOMÉTRICA E AFERIÇÃO DE AUTORIA

Uma outra demonstração que pode ser feita com o uso do R na pesquisa com dados linguísticos é a aferição da *influência do estilo de escrita* de diferentes autores. Uma análise estilométrica pode mostrar padrões de semelhança no estilo de cada autor, *podendo-se identificar alguma relação entre as escritas*. Como exemplificação, foi feita uma análise baseada na comparação de algumas obras de Franz Kafka (1883-1924) com algumas obras de Heinrich von Kleist (1777-1811). A literatura aponta praticamente como uma unanimidade que existe uma influência de von Kleist na obra de Kafka (cf. PETERS, 1966; FURST, 1985; GRANDIN, 1987; ENGELSTEIN, 2006; SHAHAR, 2007; MEHIGAN, 2011; entre outros). O objetivo foi verificar até que ponto essa influência é captada por meio de uma análise quantitativo-computacional, de forma a observar qual ou quais obras de Kafka sofreram mais influência de qual ou quais obras de von Kleist. Evidentemente, os resultados se baseiam no que o método oferece em termos de comparação de palavras. As tramas, os cenários, o perfil psicológico dos personagens etc. não podem ser acessados por meio desse método. Apenas uma análise literária qualitativa seria capaz de traçar um paralelo entre essas categorias dos estudos literários. Dessa maneira, não foi nosso objetivo traçar hipóteses substanciais a respeito do que pode ter acarretado a possibilidade de que haja ou não influência de um autor sobre o outro. O objetivo foi tão somente mostrar como o R pode ser útil para identificar semelhanças de estilo entre autores diferentes, realizando, para isso, uma análise estilométrica baseada em padrões quantitativos.

O pacote utilizado foi o mesmo que foi usado nos estudos anteriores. O método adotado denomina-se *Rolling Delta* (cf. RYBICKI *et al.*, 2014; TABATA, 2014; EDER, 2015). O objetivo principal desse método é distinguir autores de obras cuja autoria é colaborativa. Por exemplo, esse método é capaz de identificar quais pontos de uma obra colaborativa possui a marca de um ou de outro autor. Ele já foi usado na análise da obra *Roman de la Rose*, um poema francês do século XIII que foi escrito de forma colaborativa por Guillaume de Lorris e por Jean de Meun (cf. EDER, 2015). A análise revelou quais partes da obra foram escritas por Guillaume de Lorris e quais foram escritas por Jean de Meun, tomando como referência um gráfico gerado que apresenta o

número de palavras da obra de forma sequencial, indicando, aproximadamente, em qual trecho do número de palavras se encontra a colaboração de cada autor.

Esse método é usado também para outras finalidades, como a de aferir a influência do estilo de escrita de diferentes autores. O'Sullivan *et al.* (2018) analisaram as semelhanças existentes entre obras de James Joyce e de Flann O'Brien, apontando quais obras do último autor são estilisticamente mais semelhantes às obras do primeiro autor. O'Sullivan *et al.* (2018, p. 3) justificam o uso desse método dizendo que "[…] para identificar possíveis peculiaridades no desenvolvimento sequencial dos textos analisados, usamos o Rolling Delta [...], que forma uma assinatura autoral com base em um conjunto de textos e, em seguida, aplica essa impressão digital a outro texto"[11]. A justificativa dos autores conta com uma pequena explicação de como o método funciona: um conjunto de textos de referência são comparados com o texto-teste, e daí se extraem as semelhanças estilísticas com base numa análise quantitativa das palavras. No nosso caso, os textos de referência foram as obras de Kafka que, por sua vez, foram comparadas com as obras de von Kleist, que constituem os textos testes. Cada texto-teste foi comparado individualmente com os textos de referência, o que acabou gerando gráficos diferentes para cada texto-teste a ser comparado.

O que está por trás do método *Rolling Delta* envolve algumas questões relacionadas a propriedades matemáticas que não serão alvo de discussão deste trabalho. É importante mencionar, contudo, que esse método se baseia num conceito denominado *moving window*. Esse conceito é um procedimento capaz de acessar fenômenos lineares, tais quais a análise de obras literárias. Eder (2015) fornece uma explicação afirmando que o objetivo desse procedimento é medir propriedades matemáticas de um subconjunto consecutivo k de uma sequência de eventos extraídas do início de uma sequência, a partir de uma sequência de eventos, denominada uma série temporal, que consiste de N elementos. A partir daí, move-se uma 'janela' do tamanho k através de toda a série temporal até que a posição $x = N - k$ seja alcançada. Consequentemente, é possível obter informações de segmentos específicos da série temporal em seu desenvolvimento, observando-se regularidades periódicas, bem como idiossincrasias locais. O conceito de *moving window* é importante porque ele é a base de aplicação do método *Rolling Delta*. De acordo com Eder (2015, p. 2), nesse método,

[...] o procedimento de janelamento padrão é executado em todo um corpus de referência: um centroide representativo para cada texto de referência que consiste na frequência relativa média para cada uma das N palavras nas janelas extraídas do texto é calculado. Em seguida, o texto de teste também é dividido em janelas e uma medida de distância (nesse caso, Delta de Burrow) entre cada janela de texto e cada centroide de referência é calculada. Os resultados são visualizados usando um conjunto de curvas – uma para cada texto de referência. A etapa final envolve identificar, para cada janela, a linha mais baixa, ou seja, o texto de referência mais semelhante. Sempre que ocorre um takeover (cruzamento de linha), a janela respectiva é considerada para revelar uma mudança estilística[12].

Essa explicação ficará mais clara quando apresentarmos os gráficos resultantes das análises empreendidas por meio deste método. Por ora, basta saber que o objetivo do método é dividir um texto-teste em várias janelas amostrais e contrastá-las uma a uma com os textos de referência. O gráfico gerado mostra o nível de semelhança de estilo em cada texto de referência, sendo que quanto mais próximo da base maior é a semelhança estilística com o texto-teste. Abaixo, é apresentado o quadro 2 com os textos de referência que serão confrontados com os textos testes. As obras de Kafka são os textos de referência, ao passo que as obras de von Kleist são os textos testes.

[11] Tradução nossa do original: "[…] to identify possible peculiarities in sequential development of the analyzed texts, we use Rolling Delta [...], which forms an authorial signature based on one set of texts, and then applies that fingerprint to another text".

[12] Tradução nossa do original: "[…] the standard windowing procedure is run throughout a reference corpus: a representative centroid for each reference text that consists of the mean relative frequency for each of the *N* words in the windows extracted from the text is calculated. Next, the test text is also divided into windows and a distance measure (in this case, Burrow's Delta) between each text window and each reference centroid is computed. The results are visualized using a set of curves – one for each reference text. The final step involves identifying, for each window, the lowest line, i.e. the most similar reference text. Whenever a takeover (line crossing) occurs, the respective window is assumed to reveal a stylistic change".

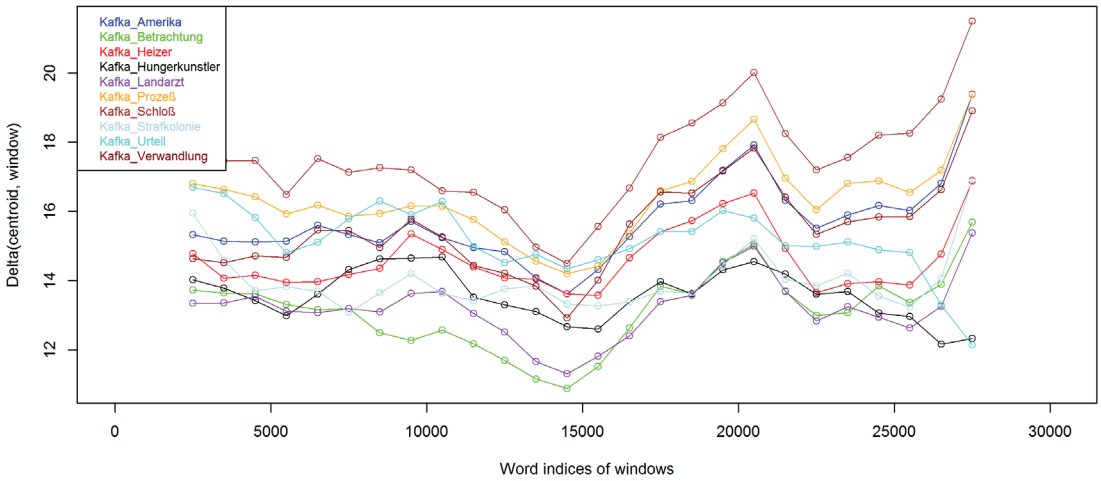
Quadro 2: Obras de Kafka e de von Kleist

Franz Kafka (1883-1924)	**Heinrich von Kleist (1777-1811)**
Betrachtung [Considerações] (1912)	*Das Käthchen von Heilbronn* (1807-1808)
Der Heizer [O Foguista] (1913)	*Penthesilea* (1808)
Das Urteil [O veredito] (1913)	*Prinz Friedrich von Homburg* (1809-1810)
Die Verwandlung [A metamorfose] (1915)	*Michael Kohlhaas* (1810)
Ein Landarzt [Um médico rural] (1919)	*Der zerbrochene Krug* (1811)
In der Strafkolonie [Na colônia penal] (1919)	
Ein Hungerkünstler [Um artista da fome] (1922) *Der Prozeß* (1925) [O processo] *Das Schloß* (1926) [O castelo] *Amerika* (1927) [O desaparecido]	

Fonte: Elaboração própria

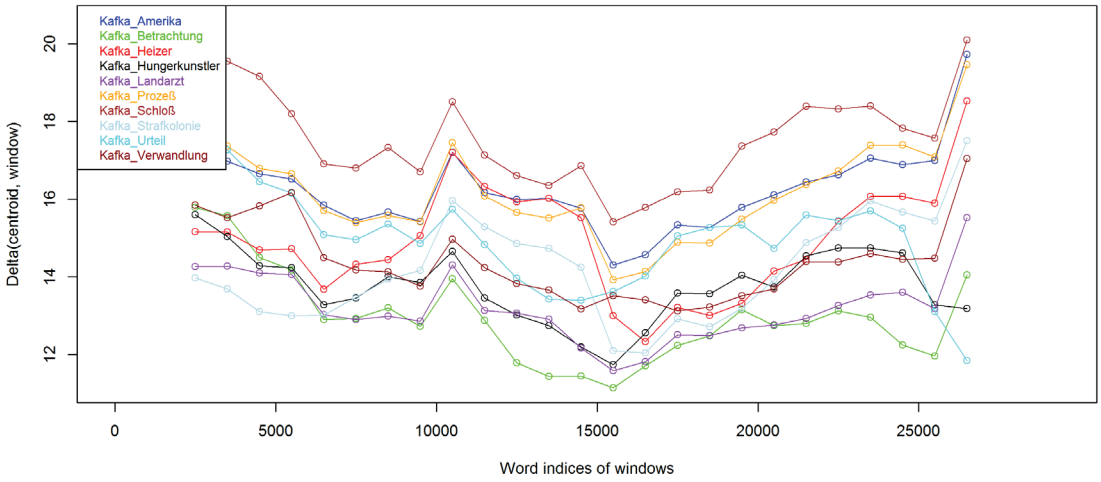
Foram dez textos de referência, portanto, o gráfico contará com dez curvas que se movem através das janelas. Como foram cinco textos testes, houve cinco gráficos distintos confrontando as dez obras de referência. O pacote comporta uma análise para a língua alemã, que deve ser selecionada durante a aplicação do método no R. O nome de cada obra de Kafka que aparece no canto superior esquerdo do gráfico é uma abreviação dos títulos originais. Abaixo, são apresentados os gráficos (do 4 ao 8) para cada análise empreendida com os textos testes, *Das Käthchen von Heilbronn, Penthesilea, Prinz Friedrich von Homburg, Michael Kohlhaas, Der zerbrochene Krug,* respectivamente.

Gráfico 6: Obras de referência vs. *Käthchen von Heilbronn*



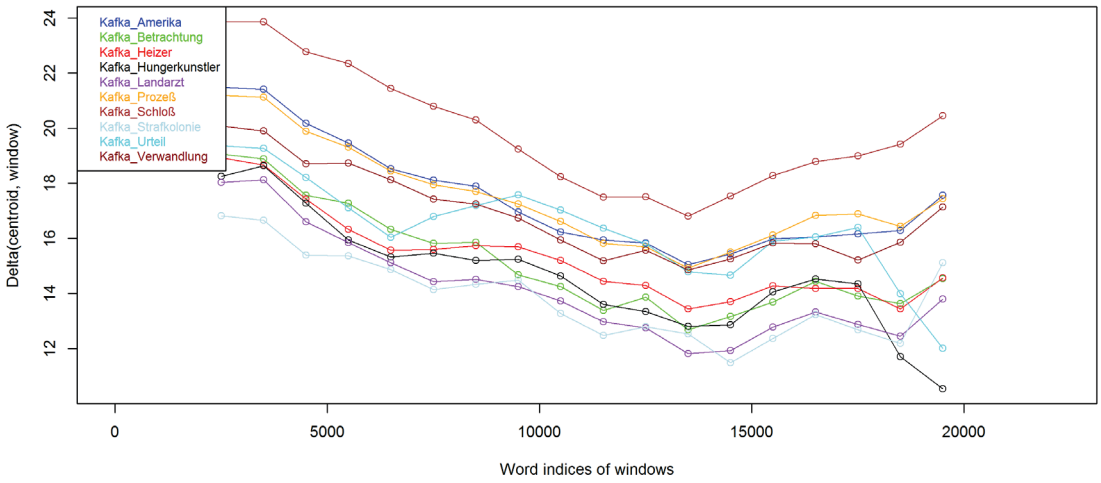
Fonte: ElAboração própria

Gráfico 7: Obras de referência vs. *Penthesilea*



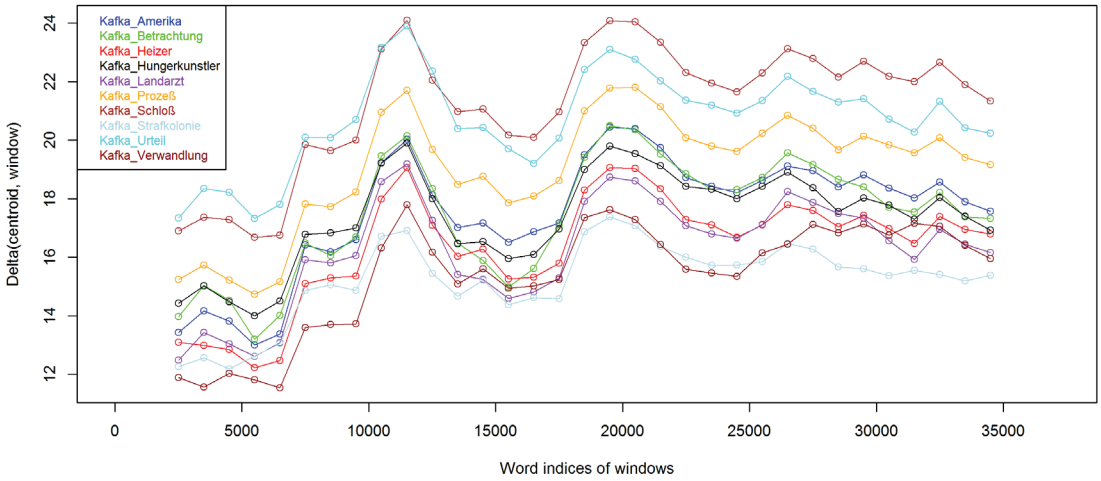
Fonte: Elaboração própria

Gráfico 8: Obras de referência vs. *Prinz Friedrich von Homburg*



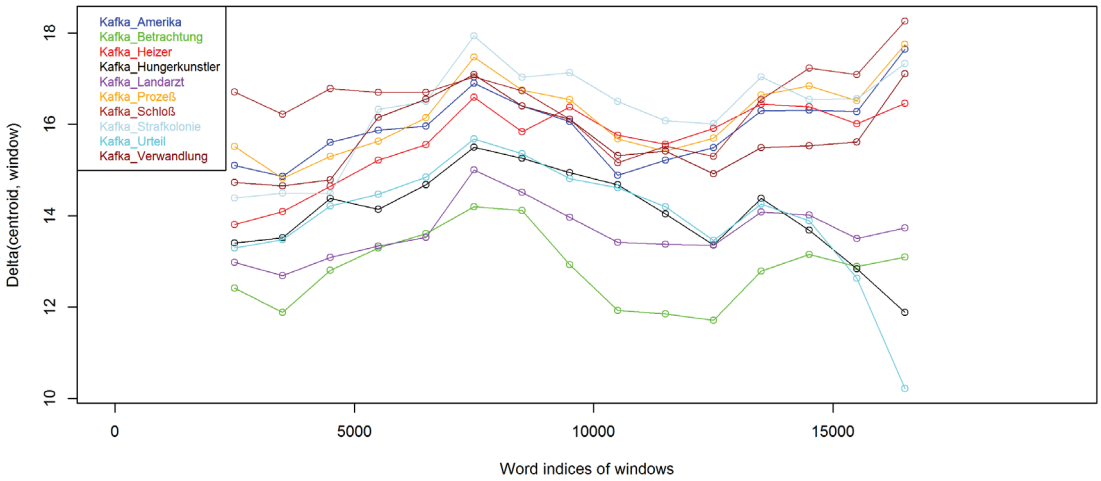
Fonte: Elaboração própria

Gráfico 9: Obras de referência vs. *Michael Kohlhaas*



Fonte: Elaboração própria

Gráfico 10: Obras de referência vs. *Der zerbrochene Krug*



Fonte: Elaboração própria

Em primeiro lugar, um comentário que pode ser feito é que, de uma forma geral, *o estilo de escrita de Kafka corresponde ao estilo de von Kleist em algumas passagens específicas, mas há outras passagens nas quais não se nota uma correspondência*. Isso pode ser observado a julgar pela distância que se encontram as curvas correspondentes às obras de referência da base do gráfico. No entanto, algumas constatações a respeito de semelhanças estilísticas podem ser feitas. Começando por *Käthchen von Heilbronn,* a obra de Kafka que parece ter sido mais influenciada pela referida obra de von Kleist é *Betrachtung* e, em menor escala, *Ein Landarzt*. Esse padrão é praticamente o mesmo encontrado em *Penthesilea*, isto é, as obras de Kafka que apresentam um estilo semelhante com *Penthesilea* é *Betrachtung* e, em menor escala, *Ein Landarzt*. No caso de *Prinz Friedrich von Homburg,* não se nota uma influência clara nas obras de Kafka. Num grau menor, poderiam ser citadas *Ein Landarzt* e *In der Strafkolonie*. Mais próximo da última janela, *Ein Hungerkünstler* apresenta uma clara semelhança, mas é um trecho muito pontual para se dizer que a obra em sua integridade foi influenciada por *Prinz Friedrich von Homburg*. Analisando *Michael Kohlhaas*, é possível notar que há semelhança de estilo com uma janela inicial *Die Verwandlung*. Posteriormente, não há uma correlação muito clara. Em menor escala, poderiam ser mencionadas a própria *Die Verwandlung* e a *In der Strafkolonie.* Por fim, *Der zerbrochene Krug* mostra uma influência em menor escala em *Betrachtung* e uma semelhança de estilo forte com *Das Urteil* apenas na última janela. Poder-se-iam resumir tais achados no quadro 3.

Quadro 3: Comparação estilística entre Kafka e von Kleist

Obras de von Kleist	Obras de Kafka com semelhança estilística mais marcada	Obras de Kafka com semelhança estilística menos marcada
Käthchen von Heilbronn	*Betrachtung*	*Ein Landarzt*
Penthesilea	*Betrachtung*	*Ein Landarzt*
Prinz Friedrich von Homburg		*Ein Landarzt, In der Strafkolonie*
Michael Kohlhaas		*Die Verwandlung, In der Strafkolonie*
Der zerbrochene Krug		*Betrachtung, Das Urteil*

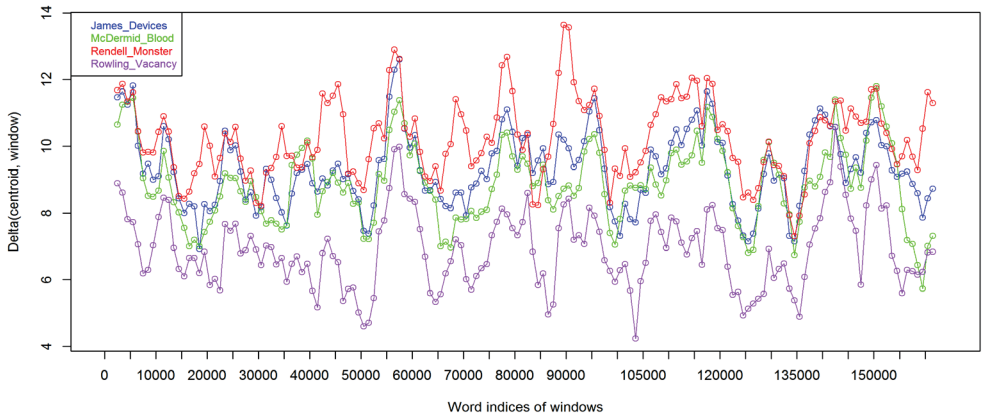
Fonte: Elaboração própria

Com a análise quantitativo-computacional empreendida por meio do *software* R, constata-se que a obra de Kafka que parece carregar mais semelhança estilística com as obras de von Kleist é a *Betrachtung*, a primeira obra de Kafka. De um ponto de vista não literário, faz sentido que essa seja a obra que tenha sido mais impactada por uma influência do estilo de escrita de von Kleist, porque, na primeira obra escrita por um autor, seu estilo ainda é embrionário, está em fase de construção uma maturidade de escrita literária maior. Um dado interessante que poderia ser acrescentado a essa análise é que, quando foi feita a publicação de sua primeira obra, isto é, *Betrachtung*, Kafka pediu ao seu editor que a publicasse com o mesmo tipo de papel em que a obra *Anekdoten* de von Kleist foi publicada (cf. PETERS, 1966). Com os resultados mostrados neste estudo, pode-se dizer que a influência que von Kleist apresenta sobre *Betrachtung* parece ser mais do que o tipo de papel em que a obra foi publicada. Contudo, apenas uma análise literária especializada e de cunho qualitativo poderia aferir as razões pelas quais há certa influência de von Kleist não apenas em *Betrachtung*, mas também em *Ein Landarzt,* em *In der Strafkolonie* e em *Die Verwandlung*.

Por fim, uma última aplicação do método *Rolling Delta* pode ser ilustrada por meio de um caso de detecção de autoria. Para exemplificar, foi analisado um caso já resolvido na literatura. Em 2013, o cientista da computação Patrick Juola recebeu um *email* de um repórter do jornal Sunday Times dizendo que ele havia conseguido uma dica de que J. K. Rowling havia escrito a obra *The Cuckoo's Calling* sob o pseudônimo de Robert Galbraith. Essa dica tinha certa plausibilidade, já que Rowling e Galbraith tinham o mesmo agente e o mesmo editor. Além disso, era a primeira obra de Galbraith, e ele sabia descrever o vestuário feminino muito bem. Contudo, faltava uma prova mais contundente. O repórter pediu a Juola para que o *software* desenvolvido pelo próprio cientista da computação fosse utilizado para resolver esse caso. Juola tomou obras de três autoras de língua inglesa que escreviam sobre temas semelhantes aos de Galbraith, bem como uma obra de Rowling para comparação com as outras três. Através dessa comparação, os resultados gerados pelo *software* indicaram que os traços da obra de Galbraith se aproximavam mais dos de Rowling do que dos traços das outras autoras. Juola diz que isso não prova efetivamente que Rowling é a autora da obra de Galbraith (cf. JUOLA, 2013a; 2013b), mas é uma indicação de uma semelhança que, somada à informação que ele recebeu do repórter, pode ser confrontada com o questionamento outrora levantado. Esses resultados foram suficientes para que o *Sunday Times* abordasse o agente de Rowling. Posteriormente, a escritora admitiu que ela é a autora da obra *The Cuckoo's Calling*, sob o pseudônimo de Robert Galbraith.

Utilizamos o *Rolling Delta* para fazer uma análise estilométrica, observando quais as obras de quatro autoras se aproximam mais do estilo de Galbraith. As autoras foram as mesmas selecionadas no estudo de Juola, representadas pelas seguintes obras: *Devices and Desires*, de P. D. James, *The Wire in the Blood*, de Val McDermid, *The Monster in the Box*, de Ruth Rendell. Embora as autoras sejam as mesmas utilizadas no estudo de Juola, as obras são diferentes, exceto a de McDermid. Além dessas três obras, também se utilizou a primeira obra de J. K. Rowling escrita após a série Harry Potter, intitulada *The Casual Vacancy*. Selecionou-se a segunda obra de Galbraith, intitulada *The Silkworm*. Vale destacar que Juola utilizou a primeira obra de Galbraith em seu estudo, *The Cuckoo's Calling*. Abaixo, pode-se observar o gráfico 9 do método *Rolling Delta* aplicado às obras supracitadas.

Gráfico 11: Obras de referência vs. *The Silkworm*



Fonte: Elaboração própria

Com a inspeção do gráfico, fica bastante nítido o fato de que a obra cujo estilo mais se aproxima ao da obra *The Silkworm* é a obra *The Casual Vacancy* de Rowling. A curva que representa essa obra é a que mais se aproxima da base, e existem, pelo menos, quatro pontos que mostram trechos de estilo muito semelhantes com os de Galbraith. Como Juola afirma, isso não prova inicialmente que Galbraith é Rowling, mas fornece fortes indícios de que existem semelhanças entre Rowling e Galbraith que denotam um estilo mais próximo um do outro do que se comparado a outras autoras que escrevem o mesmo tipo de ficção. O método é utilizado sempre como uma ferramenta que permite a aferição de uma hipótese, mas a confirmação dessa hipótese no mundo real depende de outros fatores que estão fora do alcance do método. Nesse caso, para se ter a total certeza de que Galbraith é Rowling, foi necessário que a própria autora admitisse, embora o método utilizado por Juola oferecesse sólidos indícios de uma semelhança estilística entre as obras, assim como oferece o *Rolling Delta* neste estudo.

4 CONSIDERAÇÕES FINAIS

Neste artigo, replicamos o estudo de Eder *et al*. (2016) com obras de Mark Twain, Herman Melville, Franz Kafka, Heinrich von Kleist e J. K. Rowling para demonstrar como a utilização do R pode facilitar análises de aferição de autoria a partir de dados da língua natural. Com as análises de PCA, *Cluster* e Árvore de Consenso, foi possível, por um lado, demonstrar a individuação de autoria e a identificação de obras distintas de um mesmo autor. Por outro lado, as análises com Zeta de Craig e *Rolling Delta* se mostraram relevantes para a identificação do estilo de autoria, de modo que a primeira realçou as preferências lexicais de cada autor e obra, e a segunda realçou as influências do estilo de escrita dos autores.

AGRADECIMENTOS

Os autores agradecem os comentários e as sugestões dos pareceristas anônimos que avaliaram este artigo.

REFERÊNCIAS

ARGAMON, S. Interpreting Burrow's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, v. 23, n. 2, p. 131-147, 2008.

BARTOSZUK, M.; GAGOLEWSKI, M. SimilaR: R Code Clone and Plagiarism Detection. *The R Journal*, v. 12, n. 1, p. 367-385, 2020.

BOEHMKE, B.; GREENWELL, B. *Hands-On Machine Learning with R*. New York: CRC Press, 2019.

BURROWS, J. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, v. 17, n. 3, p. 267-287, 2002.

BURROWS, J. Who wrote Shamela? Verifying the Authorship of a Parodic Text. *Literary and Linguistic Computing*, v. 20, n. 4, p. 437-450, 2005.

BURROWS, J. All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, v. 22, n. 1, p. 27-47, 2006.

CRAIG, H.; KINNEY, A. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press, 2009.

EDER, M. Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, v. 6, p. 99-114, 2011.

EDER, M. Rolling stylometry. *Digital Scholarship in the Humanities*, v. 31, n. 3, p. 1-13, 2015.

EDER, M. Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities*, v. 32, n. 1, p. 50-64, 2017.

EDER, M; RYBICKI, J.; KESTEMONT, M. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, v. 8, n. 1, p. 107-121, 2016.

ENGELSTEIN, S. The Open Wound of Beauty: Kafka Reading Kleist. *The Germanic Review: Literature, Culture, Theory*, v. 81, n. 4, p. 340-359, 2006.

FURST, L. Reading Kleist and Kafka. *The Journal of English and Germanic Philology*, v. 84, n. 3, p. 374-395, 1985.

GRANDIN, J. *Kafka's Prussian Advocate*: A Study of the Influence of Heinrich von Kleist on Franz Kafka. Columbia: Camden, 1987.

HENNIG, C.; MEILA, M.; MURTAGH, F.; ROCCI, R. *Handbook of Cluster Analysis*. Boca Raton: CRC Press, 2016.

HOOVER, D. Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style*, v. 41, n. 2, p. 174-203, 2007a.

HOOVER, D. Quantitative Analysis and Literary Studies. In: SCHREIBMAN, S.; SIEMENS, R. (ed.). *A Companion to Digital Literary Studies*. Oxford: Blackwell. 2007b. p. 517-533.

HOOVER, D. The Craig Zeta Spreadsheet. *Digital Humanities 2010* [Book of Abstracts], London: King's College London, 2010. Disponível em: http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-659.html. Acesso em: 25 ago. 2020.

JOLLIFFE, I. *Principal Components Analysis*. 2 ed. New York: Springer, 2002.

JOSHI, S. Sentiment Analysis on Whatsapp Group Chat Using R. In: SHUKLA, R.; AGRAWAL, J.; SHARMA, S.; TOMER, G. (orgs.). *Data, Engineering and Applications*. Vol. 1. Gateway East: Springer, 2019. p. 47-56.

JUOLA, P. How a Computer Program Helped Show J. K. Rowling write A Cuckoo's Calling. In: *Scientific American*, 2013a. Disponível em: https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling. Acesso em: 22 ago. 2020.

JUOLA, P. Rowling and 'Galbraith': an authorial analysis. In: *Language Log*, 2013b. Disponível em: https://languagelog.ldc.upenn.edu/nll/?p=5315. Acesso em: 22 ago. 2020.

MEHIGAN, T. The process of inferential contexts: Franz Kafka reading Heinrich von Kleist. In: MEHIGAN, T. (ed.). *Heinrich von Kleist*: Writing After Kant. Rochester: Boydell & Brewer, 2011. p. 196-226.

O'SULLIVAN, J.; BAZARNIK, K.; EDER, M.; RYBICKI, J. Measuring Joycean Influences on Flann O'Brien. *Digital Studies*, v. 8, n. 1, p. 1-25, 2018.

PENNEBAKER, J. *The Secret Life of Pronouns*: What our Words Say About Us. New York: Bloomsbury Press, 2011.

PETERS, F. Kafka and Kleist: A Literary Relationship. *Oxford German Studies*, v. 1, n. 1, p. 114-162, 1966.

QIAN, C.; HE, T.; ZHANG, R. Deep Learning based Authorship Identification. *Stanford Reports*, 2017. Disponível em: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf. Acesso em: 25 ago. 2020.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Disponível em: www.r-project.org. Acesso em: 08 jan. 2021.

RYBICKI, J.; KESTEMONT, M.; HOOVER, D. Collaborative authorship: Conrad, Ford, and rolling delta. *Literary and Linguistic Computing*, v. 29, n.3, 422-431, 2014.

SHAHAR, G. Fragments and Wounded Bodies: Kafka after Kleist. *The German Quarterly*, v. 80, n. 4, p. 449-467, 2007.

SILGE, J.; ROBINSON, D. *Text Mining with R*: A Tidy Approach. Sebastopol, CA: O'Reilly, 2017.

STACK OVERFLOW. 2020. Disponível em: https://stackoverflow.com/. Acesso em: 20 ago. 2020.

STAMATATOS, E. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, v. 60, n. 3, p. 538-556, 2009.

TABATA, T. Stylometry of collaborations: Dickens, Collins and their collaborative writings. In: *Digital Humanities 2014: Conference Abstracts*. Lausanne: EPFL-UNIL, 2014. p. 378-380.

THE COMPREHENSIVE R ARCHIVE NETWORK. 2020. Disponível em: https://cran.r-project.org/. Acesso em: 22 ago. 2020.

THE R PROJECT FOR STATISTICAL COMPUTING. 2020. Disponível em: https://www.r-project.org/. Acesso em: 20 ago. 2020.

TIDYVERSE. 2020. Disponível em: https://www.tidyverse.org/. Acesso em: 20 ago. 2020.

WIERZCHOŃ, S.; KŁOPOTEK, M. *Modern Algorithms of Cluster Analysis*. Cham: Springer, 2018.

ZHANG, C.; WU, X.; NIU, Z.; DING, W. Authorship Identification from Unstructured Texts. *Knowledge-Based Systems*, v. 66, p. 99-111, 2014.

Recebido em 12/01/2021. Aceito em 14/04/2021.