NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

O COMENTÁRIO: DA LINGUÍSTICA DO TEXTO AO *TEXT MINING*

Miguel Gonçalves de Magalhães

Tese de Doutoramento em Linguística

Janeiro, 2025

NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

O COMENTÁRIO: DA LINGUÍSTICA DO TEXTO AO *TEXT MINING*

MIGUEL GONÇALVES DE MAGALHÃES

TESE DE DOUTORAMENTO EM LINGUÍSTICA

ESPECIALIDADE EM LINGUÍSTICA DO TEXTO E DO DISCURSO

JANEIRO, 2025

DECLARAÇÕES

Declaro que esta tese é o resultado da minha investigação pessoal e independente. O seu conteúdo é original e todas as fontes consultadas estão devidamente mencionadas no texto, nas notas e na bibliografia.

Declaro ainda que tomei conhecimento do Código de Ética da Universidade NOVA de disponível na intranet e que foram respeitados os seus termos no decorrer do presente trabalho de investigação.

O candidato,

Miguel Gonçalves de Magalhães

Lisboa, 07 de janeiro de 2025

Declaro que esta tese se encontra em condições de ser apreciado pelo júri a designar.

Declaro ainda que tomei conhecimento do Código de Ética da Universidade NOVA de Lisboa disponível na intranet.

O(A) orientador(a),

J Gonçalves

Lisboa, 7 de janeiro de 2025

Tese apresentada para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Linguística, na especialidade de Linguística do Texto e do Discurso, realizada sob a orientação científica da Professora Doutora Matilde Gonçalves.

O presente trabalho é financiado por fundos nacionais portugueses, através da FCT - Fundação para a Ciência e Tecnologia, da bolsa de investigação BPD/BD/142789/2018, ao abrigo do Programa de Doutoramento FCT "KRUse – Knowledge, Representation & Use".

À Victória

AGRADECIMENTOS

Este projeto de não seria possível sem as muitas pessoas que se cruzaram comigo e que, de uma maneira ou de outra contribuíram para o seu resultado:

A **Fundação para a Ciência e para Tecnologia** pela Bolsa de Doutoramento que, através do Programa de Doutoramento "KRUse–Knowledge, Representation & Use", permitiu a minha dedicação exclusiva.

O **Centro de Linguística da Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa** pela forma como acolheu este projeto e me proporcionou as melhores condições.

A **Professora Doutora Matilde Gonçalves**, orientadora deste projeto, não só pelo acompanhamento constante, rigor científico e dedicação com que orientou este trabalho, mas também pelas palavras de apoio, incentivo, e permanente abertura de espírito às ideias.

Todos os **Professores e Investigadores** do Centro de Linguística da Universidade NOVA de Lisboa que sempre me incentivaram a prosseguir com este trabalho.

Os colegas que se cruzaram comigo nesta jornada, em especial a **Sílvia** e o **Sandro** pelo apoio e momentos de descontração, sempre necessários.

A minha **família** e amigos, sem os quais nada era possível.

O **Alexandre** pelo apoio incondicional em todos os momentos e a quem dedico este trabalho.

RESUMO

O Comentário: da linguística do texto ao *text mining*

Miguel Gonçalves de Magalhães

A presente tese desenvolve-se no quadro teórico do Interacionismo Sociodiscursivo (ISD) (Bronckart, 1997/2008) e propõe-se investigar o género textual Comentário, partindo das noções de parâmetros de género, mecanismos de realização textual e marcadores de género (Miranda, 2010), tal como foram definidos por Coutinho & Miranda (2009), como instrumentos de análise dos textos e de caracterização do comentário. O objetivo principal da investigação é compreender se a prática textual do comentário constitui um género relativamente estabilizado, com características próprias e fronteiras delimitadas em relação a outros géneros textuais, ou se se trata de um conjunto de textos sem fronteiras nítidas. Como objetivo secundário, a investigação visa identificar e sistematizar as marcas que caracterizam o género comentário Para tal, recorre-se a uma metodologia que alia a análise textual qualitativa ao uso de ferramentas de *text* e *data mining*, o que permite validar empiricamente os resultados obtidos.

A investigação desenvolve-se em três fases, cada uma dedicada a um *corpus* textual específico:

1. Análise das unidades linguísticas com o objetivo de caracterizar o comentário e avaliar se uma abordagem quantitativa é suficiente para identificar padrões que definam o género.

2. Análise dos tipos discursivos presentes nos corpora, introduzindo esta noção como uma das varáveis no contexto do *text* e *data mining.*

3. Desenvolvimento de um modelo de classificação, utilizando variáveis de nível meso (tipos discursivos) e macro (tema e atividade), sendo a variável-alvo o género textual.

A metodologia adotada privilegia a abordagem multivariada, permitindo a articulação entre os níveis micro, meso e macro de análise. A constituição e anotação de diversos corpora - incluindo textos do corpus G&T Comenta, Cetem Público e comentários jurídicos Comjur - forneceu a base empírica para a aplicação das técnicas de *data mining*. O modelo de classificação desenvolvido avalia a relevância de diferentes variáveis, demonstrando que a Atividade apresenta o maior peso na identificação do género textual, seguida pelos Tipos Discursivos (TD) e pelo Tema. Este resultado reforça a importância das práticas sociais e comunicativas no processo de análise textual, conforme sublinhado por teóricos como Bronckart (1997, na esteira de Volochinov (1929).

A análise das métricas do modelo, como a *accuracy* (91,5%), e as métricas de *precision* (94,4%) e *recall* (88,9%), evidencia a robustez e o equilíbrio do modelo desenvolvido. O uso de marcadores de género como ferramenta de anotação e análise permitiu identificar padrões que sustentam a hipótese de que o comentário pode ser considerado um género textual relativamente estabilizado, apesar das múltiplas formas que pode assumir. Além disso, a análise das previsões e dos custos associados às classificações demonstrou que o modelo é eficaz na distinção entre “Notícia” e “Comentário”, ainda que sejam necessários futuros ajustamentos para lidar com a incerteza em algumas instâncias, particularmente relacionadas com a variável Tema

Esta investigação contribui para o avanço dos estudos sobre géneros textuais ao integrar metodologias tradicionais de análise textual com técnicas inovadoras de *text mining*, oferecendo uma abordagem experimental que visa tanto caracterizar o género comentário quanto explorar o potencial das ferramentas de *data mining* no campo da linguística do texto. Assim, este trabalho preenche uma dupla lacuna: a compreensão dos padrões associados ao género textual do comentário e a criação de uma interface entre a análise textual e as metodologias de *data mining*, propondo uma via metodológica interdisciplinar que pode ser aplicada em estudos futuros.

PALAVRAS-CHAVE: Interacionismo Sociodiscursivo, Parâmetros de Género, Marcadores de Género, Tipos Discursivos, Text Mining, Análise Multivariada.

ABSTRACT

The Commentary: From Text Linguistics to Text Mining

Miguel Gonçalves de Magalhães

This thesis is developed within the theoretical framework of Sociodiscursive Interactionism (ISD) (Bronckart, 1997/2008) and aims to investigate the textual genre of Comment, drawing on the concepts of genre parameters, mechanisms of textual realization, and genre markers (Miranda, 2010), as defined by Coutinho & Miranda (2009), as tools for analyzing texts and characterizing the comment. The primary objective of the research is to understand whether the textual practice of the comment constitutes a relatively stabilized genre, with distinct characteristics and boundaries in relation to other textual genres, or whether it consists of a set of texts without clear boundaries. The secondary objective is to identify and systematize the features that characterize the comment genre. To achieve this, the study combines qualitative textual analysis with the use of text and data mining tools, allowing for the empirical validation of the results obtained.

The research is conducted in three phases, each dedicated to a specific textual corpus:

1. Analysis of linguistic units to characterize the comment and evaluate whether a quantitative approach is sufficient to identify patterns that define the genre.

2. Analysis of the discursive types present in the corpora, introducing this concept as one of the variables in the context of text and data mining.

3. Development of a classification model, using meso-level (discursive types) and macro-level (theme and activity) variables, with the target variable being the textual genre.

The adopted methodology favors a multivariate approach, enabling the articulation of micro, meso, and macro levels of analysis. The constitution and annotation of various corpora — including texts from the G&T Comenta corpus, Cetem Público, and legal comments from Comjur — provided the empirical basis for applying data mining techniques. The developed classification model evaluates the relevance of different variables, demonstrating that Activity has the greatest weight in identifying the textual genre, followed by Discursive Types (DT) and Theme. This result reinforces the importance of social and communicative practices in the textual analysis process, as emphasized by theorists such as Bronckart (1997), following Volochinov (1929).

The analysis of model metrics, such as accuracy (91.5%), and precision (94.4%) and recall (88.9%) metrics, highlights the robustness and balance of the developed model. The use of genre markers as a tool for annotation and analysis allowed for the identification of patterns that support the hypothesis that the comment can be considered a relatively stabilized textual genre, despite the multiple forms it can take. Furthermore, the analysis of predictions and associated classification costs demonstrated that the model is effective in distinguishing between "News" and "Comment," although further adjustments are necessary to address uncertainties in some instances, particularly related to the Theme variable.

This research contributes to advancing the study of textual genres by integrating traditional textual analysis methodologies with innovative text mining techniques, offering an experimental approach that aims to both characterize the comment genre and explore the potential of data mining tools in the field of text linguistics. Thus, this work fills a dual gap: understanding the patterns associated with the textual genre of comment and creating an interface between textual analysis and data mining methodologies, proposing an interdisciplinary methodological pathway that can be applied in future studies.

Keywords: Sociodiscursive Interactionism, Genre Parameters, Genre Markers, Discursive Types, Text Mining, Multivariate Analysis.

Índice

Aspetos introdutórios

A tese que agora apresento resulta do trabalho de investigação em Linguística do Texto e do Discurso (especialidade oferecida pela Faculdade de Ciências Sociais e Humanas da Universidade NOVA de Lisboa) e intitula-se "O comentário: da linguística do texto ao *text* mining". Parte da noção de *texto* enquanto produção verbal situada (Bronckart, 1997) e coloca no centro da sua análise a noção de *género* como modelo de produção e interpretação textual (Rastier, 2001; Coutinho, 2019).

Este trabalho de investigação visa responder às seguintes perguntas orientadoras e simultaneamente assumidas como objetivos:

(i) Será a prática textual do comentário um género relativamente estabilizado, com características próprias e fronteiras delimitadas relativamente a outros géneros textuais, ou estamos, pelo contrário, perante uma nebulosa de textos sem fronteiras nítidas?

(ii) Se, de facto, estamos perante um género textual estável, que características podemos elencar e como podemos validar essas características?

Para tal, constituem-se como objetivos específicos:

- realizar uma caracterização linguística e contextual das práticas textuais do comentário, alicerçada na linguística do texto e do discurso;

- convocar, sempre que necessário, outras perspetivas que, sem prejuízo, ofereçam soluções metodológicas específicas, text e *data mining* (Biber, 1995; McEnery, 2001; Santini, 2004, 2005).

Este trabalho de investigação privilegia a abordagem epistemológica desenhada no âmbito do *Interacionismo Sociodiscursivo* (Bronckart, 1997) por entender que é a perspetiva teórica e epistemológica que melhor aborda a questão das interrelações entre atividade de linguagem, género textual e texto.

O processo de investigação levado a cabo e que agora se apresenta sob a forma de tese, é orientado por uma perspetiva experimental que segue uma via de análise textual / discursiva da língua, utilizando ferramentas de *text* e *data mining*, e incide sobre os aspetos contextuais em que os textos são produzidos, articulando níveis meso e macro de análise. (Adam, 2013)

Projeto de investigação

Partindo dos pressupostos que 1) o texto depende sempre de um género ao qual pertence e 2) um texto é a materialização de um género, enquanto modelo abstrato de produção e interpretação textual, defendo que o género comentário pode ser caracterizado a partir das noções de parâmetros de género e marcadores de género (Coutinho & Miranda, 2009; Miranda, 2010). Para alcançar os objetivos, utilizarei os dados obtidos na nossa análise para executar um modelo de classificação, através de ferramentas de *data mining*.

O *corpus* recolhido no âmbito das atividades do grupo Gramática e Texto integradas no CoRus - Projeto Estratégico 2015-2020, desenvolvido pelo CLUNL mostrava que a prática textual do comentário tinha assumido um lugar de destaque nos media e nas redes sociais. Mas o sucesso da presença do comentário no espaço mediático levou a que várias e diferentes práticas discursivas e textuais fossem etiquetadas através de marcadores autorreferenciais (Coutinho & Miranda, 2009) com o termo "comentário", sem necessariamente serem comentários (podiam ser opiniões, por exemplo). E foi a partir desta constatação que a questão inicial se começou a desenhar: compreender se das diversas práticas textuais relacionadas com o comentário emerge uma nebulosa na qual coexistem um género relativamente estabilizado, ou vários géneros, ou ainda um conjunto de textos sem fronteiras delimitadas (Bronckart, 1997: 137,138). Esta reflexão constituiu-se como um dos elementos centrais desta investigação: a partir da análise de um *corpus* selecionado tentaremos perceber se é possível identificar e sistematizar as marcas de género, entendidas como elementos que caracterizam a sua identidade.

Decidimos abordar esta questão pela ótica do *text mining* e *data mining*, enquanto ferramenta metodológica de análise informática, aliando-as às noções de

texto e género do Interacionismo Sociodiscursivo (ISD). É neste aspeto que esta tese assume nitidamente um caráter experimental, ao tentar introduzir nesta área um conceito de análise do tipo mesotextual[1], a saber os tipos discursivos (Bronckart, 1997). De facto, vários projetos de *text mining* tem tentado elaborar uma metodologia que permita a identificação do género de texto, e que são apresentados em Santini (2004). O trabalho pioneiro de Biber (1989), que utilizou técnicas estatísticas multivariadas para identificar diferentes tipos de texto, resultou numa tipologia de classificação que influenciou as normas europeias para recursos linguísticos. Embora a abordagem inovadora, baseada em dados e na identificação de fatores linguísticos, a metodologia não abordou adequadamente a complexidade dos géneros mistos. Rehm (2002) focou-se na identificação de géneros em contextos específicos, propondo uma abordagem mais contextualizada. Esta metodologia trouxe uma nova perspetiva sobre a importância do contexto na identificação dos géneros textuais, mas a aplicação prática das propostas ainda é limitada e carece de validação em larga escala. Ihlstrom & Akesson (2004) exploraram a tipologia dos géneros nos jornais *online*, propondo um quadro que considera múltiplas dimensões (conteúdo, forma, funcionalidade e posicionamento). Embora bem-sucedida na abrangência da categorização dos géneros, a implementação prática e a generalização dos resultados para outros tipos de documentos ainda são desafiantes. Outros projetos como Kessler, Nunberg, & Schutze (1997) e Crowston & Kwasnik (2004) propuseram abordagens multifacetadas à classificação dos géneros, mas ambos tiveram dificuldades na sua implementação, quer pela ausência de uma aplicação prática convincente e de critérios claros para as características analisadas, quer pela dificuldade em definir características intuitivas e extraíveis a partir dos textos. Isto significa que embora haja avanços significativos na identificação automática dos géneros textuais, muitos estudos ainda enfrentam desafios em termos de aplicação prática, definição clara das categorias e para lidar com a complexidade dos textos que não se encaixam numa única classificação. É nesta linha

[1] Adam define três níveis de análise textual: o nível macro, ou seja, o texto no seu todo; o nível meso, ou seja, as unidades textuais intermédias como as sequências textuais (às quais convocamos os tipos discursivos, tal como desenvolvido por Bronckart (1997); o nível micro, ou seja, as palavras e/ou signos linguísticos (Adam, 2013: 29)

de ação que Santini (2004) sublinha a necessidade de abordagens mais flexíveis e dinâmicas nos trabalhos atuais.

Para tentarmos responder a esta abordagem mais flexível e dinâmica, vamos adotar uma abordagem dialética, ao permitir um movimento de análise que possibilite a articulação entre os elementos micro, meso e macro dos textos. A nossa metodologia de análise segue a proposta de Bronckart (1988, 1997 e 2008) para a análise dos tipos discursivos que parte de uma análise empírica de um *corpus* de textos. O autor foca-se inicialmente na identificação das unidades linguísticas e das suas possíveis interdependências, sem recorrer a parâmetros externos durante a análise textual, mas apenas posteriormente para comparar os resultados. O princípio de indissociabilidade entre significante e significado orienta o processo, que busca identificar configurações de unidades linguísticas interdependentes, ligadas a processos que são representados no texto. A metodologia segue uma abordagem de "explicação por modelos" de Piaget (1950), que consiste em identificar codependências estatísticas significativas entre marcas linguísticas, elaborar hipóteses sobre os processos subjacentes e testar essas hipóteses com dados textuais adicionais, incluindo meios de comparações interlinguísticas.

Além disso, Bronckart difere de outros autores, como Malrieu & Rastier (2001), ao delimitar previamente os segmentos textuais em tipos discursivos. A análise das propriedades linguísticas é feita separadamente para cada tipo discursivo, o que permite uma abordagem mais precisa, evitando a circularidade de analisar os segmentos a partir das propriedades linguísticas já identificadas. A metodologia procura estudar as interações entre as unidades linguísticas e os tipos discursivos de forma estruturada e sistemática, ao permitir uma análise detalhada das características linguísticas que tornam possível identificar cada tipo de discurso no nosso *corpus*.

Assim, vamos considerar que as unidades linguísticas que permitem identificar os tipos discursivos constituem o nível micro – porque corresponde à análise das variações microlinguísticas – e que os tipos discursivos, por sua vez, constituem um nível intermédio (meso) de análise. O recurso aos parâmetros externos será explorado no

nível macro, onde serão analisados as variáveis Tema e Atividade em que foram produzidos os textos, e que introduzem uma dimensão pragmática à nossa análise.

Em termos estruturais, a tese desenvolve-se em seis capítulos que passamos a explicar.

No capítulo introdutório da tese, será feito o enquadramento teórico, tendo em conta aspetos teórico-epistemológicos e definindo os conceitos-chave que irão conduzir a investigação.

No capítulo 2 será feita uma caracterização diacrónica do comentário, tentando concomitantemente fazer uma revisão da literatura sobre o comentário e de que modo esta prática textual evoluiu e que circunstâncias contribuíram para a sua estabilização (ou não). No capítulo 3 será feita uma caracterização sincrónica do comentário contemporâneo, sublinhando a sua importância no espaço mediático e a caracterização das múltiplas formas que assume. A metodologia, que ocupa o capítulo 4, tem um tratamento separado pela complexidade que a nossa análise adquire: não só pretendemos fazer uma análise textual do comentário, como pretendemos validar os resultados através de ferramentas de *text* e *data mining*. Pretendemos apresentar neste capítulo não só o *corpus* analisado como também todo o processo de análise quantitativa e qualitativa.

No capítulo 5 será feita a análise dos dados recolhidos. A análise que pretendemos mostrar neste capítulo será uma análise quantitativa e qualitativa, tentando fazer uma interpretação estatística dentro do quadro teórico do ISD, à semelhança da que Bronckart (1989) efetuou. Tentará, também, mostrar como outras ferramentas, como processos de *cluster* e gráficos de distribuição, podem auxiliar o processo de análise de texto.

O capítulo 6 é dedicado à análise do modelo de classificação obtido pelas ferramentas de *text* e *data mining*. O capítulo divide-se entre a análise das métricas, como forma de avaliação do modelo e uma reflexão dos resultados obtidos.

Na última parte da tese (que corresponde às reflexões conclusivas) serão sistematizados os principais aspetos da investigação apresentada onde se defende o uso

das ferramentas de *text* e *data mining* como um auxiliar da análise de texto, enquanto propomos algumas pistas de investigação com vista a otimizar este processo no futuro.

Pretendemos, assim, desenvolver um trabalho que preencha duas lacunas no panorama da análise de texto e do *data mining*: por um lado perceber se as práticas textuais do comentário contém padrões que nos permitam afirmar que é um género textual e, por outro lado, estabelecer uma interface entre análise de texto, na linha da linguística do texto e do discurso, e *data mining*, procurando perceber se as ferramentas de *data mining* são aplicáveis à análise do texto, ao mesmo tempo que ensaiamos a introdução do quadro teórico do ISD no *data mining.*

1. Enquadramento teórico

O trabalho que agora apresento parte das noções de texto e género. A primeira enquanto, unidade de produção verbal, oral ou escrita, que se caracteriza pela sua capacidade de organizar linguisticamente uma mensagem e garantir a coerência interna dessa mensagem (Bronckart, 1997: 73-74). Estruturados conforme regras composicionais, os textos mostram uma interdependência com o contexto social em que são produzidos, e são reconhecidos como uma unidade comunicativa superior. O texto é, deste modo, uma produção real / efetiva dentro de um contexto (Bronckart, 2008: 10). A segunda noção, género de texto, define-se como um modelo abstrato para produzir e interpretar textos que, no contexto das diferentes atividades em que são produzidos, se estabilizam de acordo com a variação sincrónica e diacrónica a que estas atividades de linguagem estão sujeitas (Coutinho, 2019: 32).

Estas definições levantam algumas questões, das quais destacamos duas: (i) a forma como estas noções se articulam com a noção de atividade, e (ii) de que forma se correlacionam e impactam mutuamente. Será na esteira destas noções e questões que nos posicionaremos na secção seguinte.

1.1. Da atividade ao texto e do género ao texto

A noção de atividade, dentro do ISD, refere-se à organização estabelecida entre os seres vivos, a partir da qual são capazes de aceder ao meio circundante e de construir elementos de representação internos (Bronckart, 1997: 30). Como Bronckart explica, esta noção de atividade parte da formulação feita por Vygotsky (1997), que por sua vez se desenvolve a partir de Marx e Engels.

Para Marx (2015), a atividade humana é fundamentalmente uma atividade prática que ocorre no mundo material. Marx considera que a atividade é a base da transformação social e da produção material, sendo o trabalho um dos principais elementos que definem a natureza humana. Argumenta, ainda, que o trabalho, como atividade, é a forma pela qual os seres humanos se relacionam com a natureza e entre si, moldando e transformando a realidade à sua volta. Deste modo, a atividade é entendida como uma prática social, isto é, uma prática coletiva que é determinada pelas condições materiais e pelas relações sociais de produção.

Vygotsky (1997) tem uma abordagem diferente, com o foco no desenvolvimento cognitivo e na interação social. Para Vygotsky, a atividade humana é mediada por signos e ferramentas culturais, como a linguagem, que desempenham um papel central no desenvolvimento mental. O autor centra-se na ideia de que as funções cognitivas superiores, como pensamento e raciocínio, se desenvolvem a partir das interações sociais e culturais, mediadas por instrumentos simbólicos. A atividade, para Vygotsky, é uma prática socialmente mediada, e a sua teoria enfatiza a interdependência entre o indivíduo e o contexto social-cultural na formação do pensamento e do conhecimento.

No quadro do ISD, Bronckart (1997: 30-31) começa por formular a possibilidade de identificar vários tipos de atividade noutras espécies animais, e que estão associadas a funções específicas como a alimentação, a reprodução ou a identificação de perigo. São atividades ligadas à sobrevivência das espécies e que, por isso, estão condicionadas a formas específicas de manifestação, que estão ligadas a processos de cooperação entre os indivíduos: *les espèces animales témoignent donc d'une activité elle-même nécessairement collective ou «sociale», en un premier sens (trop large à nos yeux) de ce terme* (Bronckart, 1997: 31). Nos animais, as atividades estão circunscritas a funções específicas, desencadeadas por um estímulo, que se traduz numa reação não-negociada. Mas o ser humano distingue-se pela complexidade que estas atividades sociais assumiram, tornando-se mesmo numa especificidade humana (Bronckart, 2008: 14). Só através da singularidade das capacidades comportamentais dos humanos, foi possível a complexificação dessas atividades coletivas, e a linguagem é o mecanismo que possibilita a organização dessas atividades, funcionando como um sistema de sinais organizados em textos.

A especificidade humana também se reflete na diversidade das atividades práticas e nas linguagens utilizadas, que não têm paralelo direto nos comportamentos de outras espécies. Os comportamentos variam conforme o ambiente, com diferentes grupos humanos a desenvolver formas de atividades práticas e linguísticas adaptadas aos seus contextos específicos. Estas, por sua vez, levam ao surgimento de subculturas e sistemas semióticos próprios, que, ao serem aplicados em atividades práticas e linguísticas, geram diversidade cultural. A historicidade e a culturalidade são aspetos

fundamentais neste processo, pois as atividades e linguagens humanas são moldadas por marcas do passado e a sua interpretação.

A complexificação das atividades humanas, juntamente com a necessidade de um entendimento / colaboração entre aqueles que participam nessas atividades, permitiu, também, o estabelecimento de formas comuns de representação do meio circundante. A necessidade de entendimento sobre o contexto da atividade e o papel dos indivíduos no processo emerge, criando uma "pretensão de designação": *les productions sonores originelles auraient été motivées par cette nécessité d'entente; d'abord temporellement et déictiquement associées à des interventions sur les objets* (Bronckart, 1997: 32). Esta necessidade de entendimento leva à criação de sinais para indicar essas intervenções. Sinais que, embora inicialmente contestados pelos outros membros do grupo, acabam por se estabilizar como uma linguagem partilhada, na qual as representações sonoras se associam a representações de aspetos do meio. A linguagem surge, neste momento, não apenas como uma ferramenta de comunicação, mas como um mecanismo de negociação social, transformando as representações individuais em representações comuns e compartilháveis. Perspetiva que converge com a noção de agir comunicativo de Habermas (1987), segundo a qual as ações são orientadas para o entendimento, onde os participantes procuram alcançar um consenso sobre o significado e a validade das suas expressões. Mediado pela linguagem, este processo permite que os indivíduos compartilhem significados e coordenem ações de forma cooperativa.

O ISD, na continuidade do trabalho de Habermas, adota a noção de mundos representados, definidos por representações coletivas, e que se estruturam em três dimensões: o mundo objetivo (relacionado com o ambiente físico), o mundo social (relacionado com a organização das tarefas e cooperação), e o mundo subjetivo (relacionado com as características pessoais dos indivíduos). Assim, a atividade da linguagem é o resultado de uma produção interativa associada às atividades sociais, e que apenas são validadas dentro do contexto em que essa atividade decorre. A atividade da linguagem é fundamentalmente uma prática humana que se articula com outras formas de atividades práticas, e deve ser entendida num contexto social complexo, onde a linguagem desempenha um papel fundamental na comunicação e interação entre os

indivíduos. Bronckart (2008) considera que a atividade da linguagem é constituída por três particularidades:

(i) A interação com outras atividades humanas: a prática da linguagem está intimamente ligada a outras formas de atividades práticas, funcionando como um mecanismo de entendimento entre os indivíduos envolvidos, e que inclui a regulação, a avaliação e a codificação das atividades práticas.

(ii) O contexto situacional: a atividade da linguagem ocorre em circunstâncias específicas, onde os humanos produzem linguagem com base em objetivos determinados e em resposta a situações concretas.

(iii) A diversidade e heterogeneidade: As atividades linguísticas são diversas e heterogéneas, o que torna difícil classificá-las de maneira rígida. Para Bronckart, embora seja possível pensar em tipologia de atividades, a complexidade e a variedade das ocorrências tornam qualquer tarefa de classificação ilusória.

Como definiu Jorge (2014) existe uma distinção entre atividade e ação da linguagem: a atividade da linguagem decorre da interação social, sendo um processo coletivo, enquanto a ação da linguagem está relacionada com o agir individual, ou seja, a intervenção pessoal do sujeito sobre a realidade. A atividade da linguagem surge da soma das ações individuais que confluem na interação social, e, por sua vez, a forma como essas interações sociais são organizadas influencia as ações individuais. Há, portanto, uma relação de interdependência e influência mútua entre a ação da linguagem, ao nível individual, e a atividade da linguagem, ao nível coletivo.

A linguagem apresenta-se como um produto interativo, associada às atividades sociais (Bronckart, 1997: 34) e que, em confronto com o valor ilocutório das formas, estabiliza o signo enquanto forma partilhada entre os actantes:

> *Mais cette analyse fait simultanément apparaître une dimension de* rupture*, qui tient au fait que les processus hérités s'appliquent désormais non plus seulement à objets physiques comme dans le monde animal, mais aussi à des* objets sociaux*, (...) qui sont conventionnellement associés à des dimensions de l'activité humaine.*
> Bronckart (2010: 22)

Isto significa, portanto, que se estabelece uma distância (rutura) entre o meio ambiente e o indivíduo, que permite uma autonomia das produções semióticas, e possibilitam a sua reorganização noutras atividades.

Partindo desta rutura entre o meio ambiente e o indivíduo, Bronckart (2008) propõe três registos (ou esferas) que intervêm na construção do significado e na organização do conteúdo textual:

(i) O mundo comum que corresponde ao mundo do produtor do texto, e que se refere às representações pessoais do autor. Este registo influencia o que ele escolhe representar no texto e como o faz, considerando o seu contexto de ação linguística (por exemplo, o modo como ele utiliza a linguagem e a forma como percebe a situação comunicativa).

(ii) Os mundos formais que são os sistemas de conhecimento coletivos organizados segundo as convenções do grupo ao qual o produtor de texto pertence, como uma cultura ou uma comunidade, que fornece as regras e as normas para a organização do conteúdo.

(iii) Os mundos discursivos, que organizam as interações entre representações pessoais e coletivas e permitem que as diferentes formas de significado se comuniquem e se relacionem dentro do texto; unidades psicológicas enformadas no processo de textualização (ou fazer texto).

Bronckart sugere que, para compreender a formação dos mundos discursivos, nos devemo focar nas ruturas pessoais (que envolvem as representações individuais do autor) e nas ruturas temporais (que dizem respeito à forma como o tempo é estruturado no discurso):

> *Selon nous, la constitution des différents mondes discursifs peut être appréhendée en s'en tenant à la prise en compte des seules ruptures "personnelle" et "temporelle", la rupture "spatiale" ne semblant sémiotisée dans la textualité que de manière connexe et secondaire (en tout cas dans les langues que nous avons étudiées), et la rupture "modale" nous paraissant avoir un tout autre statut.* (Bronckart, 2008: 63)

Estas ruturas serão abordadas com mais detalhe na secção **1.2.5** deste trabalho, quando nos debruçarmos sobre os tipos discursivos. No entanto, foram aqui convocadas porque é a partir destas ruturas, capacitadas pela linguagem, que as representações sobre o mundo se tornam autónomas. Ao longo do tempo, os signos podem ser organizados de maneira mais complexa, criando uma atividade linguística específica que vai além da comunicação do quotidiano, e se organiza em textos[2] que, pela interação com as atividades não linguísticas se diversificam em géneros (Bronckart, 1997: 35). Significa, portanto, que existe uma relação, que não se pode dissociar, entre o contexto em que são produzidos os textos e as unidades linguísticas de cada língua natural.

A diversidade dos géneros de texto é uma questão que coloca o ângulo da análise textual sobre o estudo do texto, enquanto objeto complexo (Coutinho, 1999), e que se constitui como o objeto de uma área de investigação interdisciplinar e multifacetada. Nesta perspetiva, o texto é entendido como um objeto natural com a complexidade que lhe é inerente, ao contrário dos "objetos forjados laboratorialmente ou por abstração teórica" (Coutinho, 2006: 2). Como vimos anteriormente, a ação da linguagem manifesta-se através dos textos que são a face visível dessa ação. O texto é a manifestação empiricamente atestadas das "ações humanas da ordem da linguagem" (Coutinho, 2019: 39), o que significa, por isso, que os textos são a face visível da produção linguística, e que se constituem como objetos empíricos disponíveis ao linguista.

> *La sémiotisation donne ainsi naissance à une activité proprement langagière qui s'organise en discours ou en textes. Et sous l'effet de la diversification des activités non langagières avec lesquelles ils sont en interaction, ces textes se diversifient eux-mêmes en genres.* (Bronckart, 1997: 35)

[2] A distinção entre texto e discurso tem sido objeto de várias análises, de acordo com os quadros teóricos que as enformam, e ambos os conceitos são, em alguns casos, usados como sinónimos. Bonilla (1997) citado por Miranda (2010: 61). Contrariamente a esta abordagem, Bronckart (2008: 44) faz corresponder a noção de discurso à noção de atividade linguageira (*activités langagières*): *(...) nous réservons la notion de «genre» aux seuls textes («genres de textes») et que nous proposons d'utiliser, pour les autres niveaux, les formules «sortes d'activités générales» et «sortes de discours» (ou «sortes d'activités langagières»)* (Bronckart, 2004: 102).

A partir desta afirmação, ficam evidentes os seguintes pressupostos, que subscrevemos para este trabalho:

(i) a noção de texto enquanto objeto empírico,

(ii) a relação de interdependência do texto com o contexto social de ação,

(iii) o texto enquanto representante empírico das atividades em que é produzido,

(iv) cada atividade social, na sua diversidade, organiza/regula esses mesmos textos, de acordo com as situações práticas em que circula.

1.2. Tipos e géneros de texto

1.2.1. Tipos e géneros no âmbito do ISD.

Como referimos na secção anterior, no âmbito do ISD, os textos diversificam-se em géneros dentro das atividades de linguagem e atividades sociais em que são produzidos. Mas fora do âmbito da linguística, embora a noção de género esteja difundida, a sua noção e terminologia não são consensuais, ocorrendo confusões e sobreposições entre o uso do termo "género" e "tipo" de texto, sobretudo nos autores que, trabalhando com tratamento automático de textos, pretendem desenvolver metodologias que permitam identificar e classificar textos através do género a que pertencem. Alguns exemplos destas confusões e sobreposições podem ser encontradas em trabalhos como Kessler, Nunberg, & Schutze (1997)[3] e Finn & Kushmerick (2006)[4].

Nesta secção, vamos começar por rever a posição de Biber (1995) e Santini (2004, 2005), que ensaiaram uma distinção entre tipo e género de texto, para, posteriormente, enquadrar estes conceitos na perspetiva do ISD. Estes dois autores trabalharam na identificação e classificação automática de textos. É importante referir que, embora tenham dedicado uma parte do seu trabalho à distinção entre tipo e género, estes autores não fizeram uma reflexão teórica sobre estas noções. Os seus trabalhos são aqui

[3] Os autores usam o termo género para designar um conjunto de textos pertencente a uma *classe* de textos com um objetivo comunicativo ou outros traços funcionais (Kessler et al., 1997: 33).

[4] Os autores usam indistintamente os termos *estilo* e *género*, e definem género como uma *classe* de textos que têm um *tipo* semelhante (Finn & Kushmerick, 2006: 1506).

convocados por serem pioneiros na preocupação da análise linguística no tratamento automático de textos e, mais especificamente, na identificação e classificação automática de géneros textuais. Ambos se fundamentam na ideia de que o critério mais objetivo para identificar e classificar os géneros textuais reside na análise das unidades e das regras linguísticas específicas que estes mobilizam, embora este critério comporte a dificuldade que advém do volume de informação que implica ser analisada e codificada. Na segunda parte desta secção, vamos rever a posição de alguns autores que têm refletido teoricamente sobre as noções de género e tipo de texto, com especial foco sobre aqueles que contribuíram para a noção de género e tipo de texto dentro do ISD.

Tanto Biber (1995) como Santini (2004) têm argumentado a distinção entre tipo e género, sobretudo no tratamento automático de textos. Biber (1995: 70) considera que os tipos de texto são conjuntos de textos que se assemelham linguisticamente, enquanto o género[5] é uma categorização baseada em critérios externos. Dando como exemplo os textos de ficção científica, o autor considera que um texto pertence ao género ficção científica pela intenção autoral, mas que linguisticamente pode ser semelhante a textos científicos do tipo académico. A dimensão linguística da análise é estabelecida através de coocorrência de padrões nos textos, que são agrupados e interpretados em termos funcionais (Biber, 1995: 13). Esta oposição entre elementos externos e internos, embora promissora para o tratamento automático de textos, tem algumas lacunas. Como aponta Lee (2001: 40), há várias preocupações em relação à abordagem de Biber, das quais destacamos duas: a primeira é que a tipologia textual desenhada por Biber (1989) coloca a classificação ao nível dos textos individuais e não agrupados em géneros, significando que conjuntos de textos que, pelas suas características externas, poderiam ser agrupados sob o mesmo género, podem ser classificados em diferentes tipos de texto por causa das diferenças linguísticas. A segunda é saber se as dimensões internas que podem definir um critério tipológico é efetivamente possível de ser encontrado, e se essas dimensões são estáveis e úteis.

Esta posição relativamente à noção de tipo de texto distancia-se de conceitos formulados anteriormente (Beaugrande & Dressler, 1981; Werlich, 1976), em que o tipo

[5] registo depois de 1995 (Biber, 1995a).

de texto estava associado às categorias retóricas tradicionais - narrativo, descritivo, expositivo e argumentativo, ou outras divisões em número variável (Cf. Santini (2004: 4)), colocando o foco da análise na função comunicativa do texto. Ao contrário de Biber (1989) que desenvolveu uma análise multidimensional de caráter indutivo, Santini (2004, 2005) tenta combinar uma análise indutiva com uma análise dedutiva. A abordagem é dedutiva porque parte de um número limitado de tipos de textos (derivados das categorias retóricas tradicionais) e é indutiva porque o processo inferencial é baseado em *corpora*, através do cálculo do valor da probabilidade para uma hipótese (Santini, 2006: 70). Como podemos observar, a autora também coloca o foco da análise sobre a função comunicativa, mas define o tipo de texto em relação às estratégias retóricas / discursivas realizadas no próprio texto, enquanto o género é definido como as convenções socioculturais, textuais e linguísticas aplicadas no texto (Santini, 2005: 3). A autora propõe (2005: 3) uma metodologia de análise que permite determinar o contexto comunicativo a partir de pistas linguísticas internas ao texto e dá como exemplo o artigo académico, uma composição textual amplamente aceite e legitimada pelos agentes que representam o contexto sociocultural em que o género circula. A composição exibe características específicas em termos de organização estrutural e uso linguístico, primordialmente orientada para uma finalidade argumentativa que estabelece a sua tipologia textual.

Em resumo, as noções de tipo e género de texto têm evoluído significativamente, especialmente no contexto do tratamento automático de texto. Os estudos mais recentes, que se baseiam na análise de *corpora*, destacam a importância de distinguir claramente entre tipo e género de texto. Além disso, as abordagens utilizadas para a identificação e classificação automática de textos, quer por tipo, quer por género, fundamentam-se numa oposição entre critérios internos, de natureza linguística, e critérios externos, socioculturais.

1.2.2. Tipos e géneros no âmbito do ISD

Como referimos na secção anterior, as noções de tipo e género textual não foram alvo de reflexões teóricas por parte daqueles que desenharam uma possível distinção dos mesmos, dentro da atividade de classificação automática de textos. Como adverte Santini:

> *It is worth noting that most projects on automatic genre identification/classification do not bother very much with these theoretical issues: in many cases what they aim to achieve is a classification of documents not based on the "content", but on other features. (Santini, 2004: 3)*

Mas dentro do trabalho que nos propomos fazer, é importante que se faça esta reflexão sobre a diferença entre tipos e géneros textuais para que possamos encontrar ferramentas adequadas que permitam abordar a questão do género. Esta reflexão irá circunscrever-se às noções que consideramos terem mais impacto na análise do texto.

Adam (2001) sustenta ser impossível definir tipologias ao nível do texto, uma vez que ele é demasiado complexo e heterogéneo para manifestar regularidades linguísticas. Propôs situar as regularidades "relato", "descrição", "argumentação", "explicação" e "diálogo" a um nível chamado "sequencial" e definiu as sequências como unidades composicionais superiores à frase-período, mas muito inferiores à unidade global texto. No seu modelo dos planos de organização textual (Adam, 1992, 2001), aquilo que é apelidado de tipologia textual é uma dominância que ocorre ao nível da estrutura composicional[6] dos textos, e que Adam (1992: 20) faz corresponder à noção de sequência. O autor coloca a organização interna dos textos no centro de uma análise ascendente dos textos (as unidades mais simples organizam-se em frases ou proposições que, por sua vez, se organizam em unidades mais complexas, denominadas por sequências).

Para (Adam, 2012: 194) os tipos de texto são, por isso, derivados das sequências textuais, que ocorrem no nível 5 (N5) , a saber a estrutura composicional, constituída pelas sequências textuais e os planos de texto (cf. Figura 1).

[6] O autor identifica os tipos de texto com a textualidade (Adam, 2001: 11).

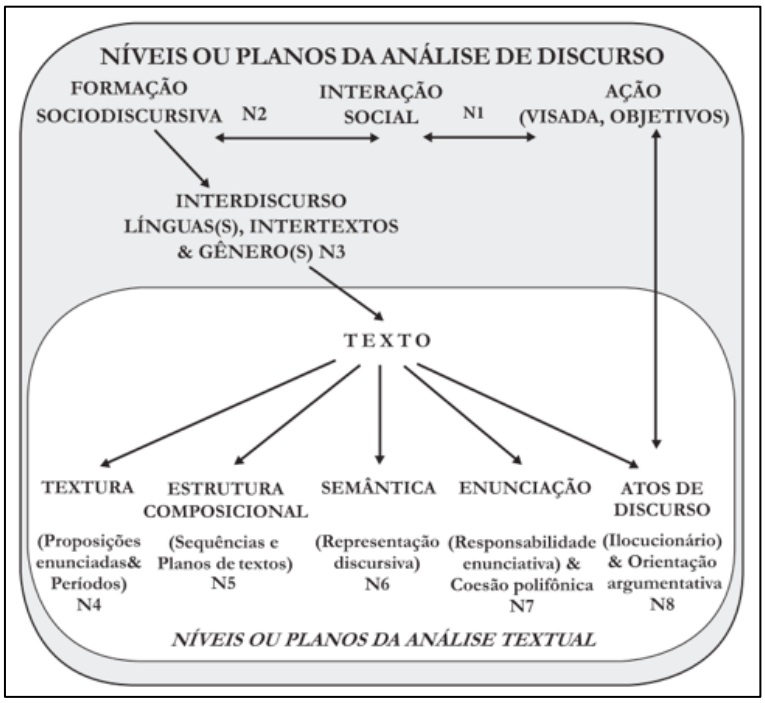


Figura 1:Níveis ou planos de análise, extraído de Adam (2012: 193).

As sequências que fazem parte de N5 são, para Adam (2008: 204), unidades textuais complexas que estabelecem relações hierárquicas entre si, e que ocupam lugares específicos na organização do texto, sendo constituídas por proposições "pré-formatadas" de sequência. As diferentes combinações destas proposições em sequências correspondem a cinco tipos de relações (denominadas de *narrativa*, *argumentativa*, *explicativa*, *dialogal* e *descritiva*), que são culturalmente reconhecidas, e que permitem reconhecer e organizar a informação textual.

Como podemos observar pelo esquema da Figura 1, o modelo de Adam baseia-se num princípio de composicionalidade, que foi criticado por outros autores (Bronckart, 2008; Rastier, 2001). Para Adam (1992: 16) a complexidade dos textos só pode ser abordada tipologicamente de um ponto de vista modular porque a heterogeneidade das produções linguísticas impede as abordagens tipológicas.

1.2.3. Análise de texto e análise de género

A análise de texto e de género coloca vários problemas que têm sido abordados em vários trabalhos (Coutinho & Miranda, 2009; Gonçalves & Miranda, 2008; Miranda, 2010). O principal obstáculo à análise de textos e de géneros é o facto de terem naturezas distintas, mas de estarem relacionados, e de não poderem ser analisados

independentemente: o género tem uma natureza abstrata e heterogénea, enquanto o texto é um objeto empírico. De acordo com Gonçalves & Miranda (2008), a análise textual incide sobre os textos empíricos, e pretende dar conta das especificidades desses textos. Por outro lado, a análise de género procura identificar os traços específicos de um determinado género, os quais só podem ser acedidos através da análise de textos empíricos. Deste modo, a análise de texto é uma etapa para aceder ao género de texto, mas que deve ter em conta o género de que relevam, e que Coutinho & Miranda (2009: 35) sintetizaram na expressão "duality of genre and text".

As ferramentas para analisar texto e género têm de refletir esta dualidade, e explicitar quais os dados que pertencem a cada um dos domínios e de que forma se relacionam entre si. Assim, no trabalho de análise que propomos realizar na presente tese, serão integradas as noções de parâmetros de género, mecanismos de realização textual e marcadores de género (Coutinho & Miranda, 2009; Miranda, 2010). Como referimos anteriormente, no âmbito do ISD, os textos diversificam-se em géneros dentro das atividades de linguagem em que são produzidos, significando, por isso, que os textos correspondem a ações de linguagem. Para a concretização desta produção, são mobilizadas as representações que o sujeito tem (sincrónica e diacronicamente) do contexto e dos diferentes géneros que são elaborados pela ação de linguagem, e que funcionam como modelos disponíveis. Estes "modelos" são representativos da identidade de um género e são denominados de parâmetros genéricos.

> *Qualquer texto se inscreve assim num género, que a avaliação da situação retém como mais adequado, de entre o conjunto de géneros disponíveis(...).* (Coutinho, 2019: 39)

No momento em que o sujeito fixa esses parâmetros no texto empírico, ocorre uma gestão dos recursos semiolinguísticos, que atualizam os parâmetros, e que são os mecanismos de realização textual. Neste sentido, é nesta atualização que reside a singularidade de cada texto porque não existe uma relação biunívoca entre parâmetros genéricos e mecanismos de realização textual: o mesmo parâmetro pode ser realizado através de diferentes mecanismos linguísticos. A introdução das noções de parâmetros genéricos e mecanismos de realização textual evidencia a necessidade de partir da análise de textos empíricos, e conduz-nos ao momento da receção dos

textos, em que os mecanismos de realização textual, ao indiciarem os parâmetros, desempenham a função de marcadores de género. Os marcadores de género são, portanto, o modo como o sujeito reconhece a identidade de género, e funciona ao nível da interpretação textual.

Na Figura 2 podemos observar os processos que estão subjacentes às relações entre parâmetros genéricos, mecanismos de realização textual e marcadores de género, representados por setas, e observar como este modelo de análise permite articular a descrição do texto com a descrição do género. O facto de o processo de atualização dos marcadores de género em parâmetros genéricos ser graficamente representado de modo diferente tem como objetivo destacar o valor distinto do parâmetro que está a ser atualizado.

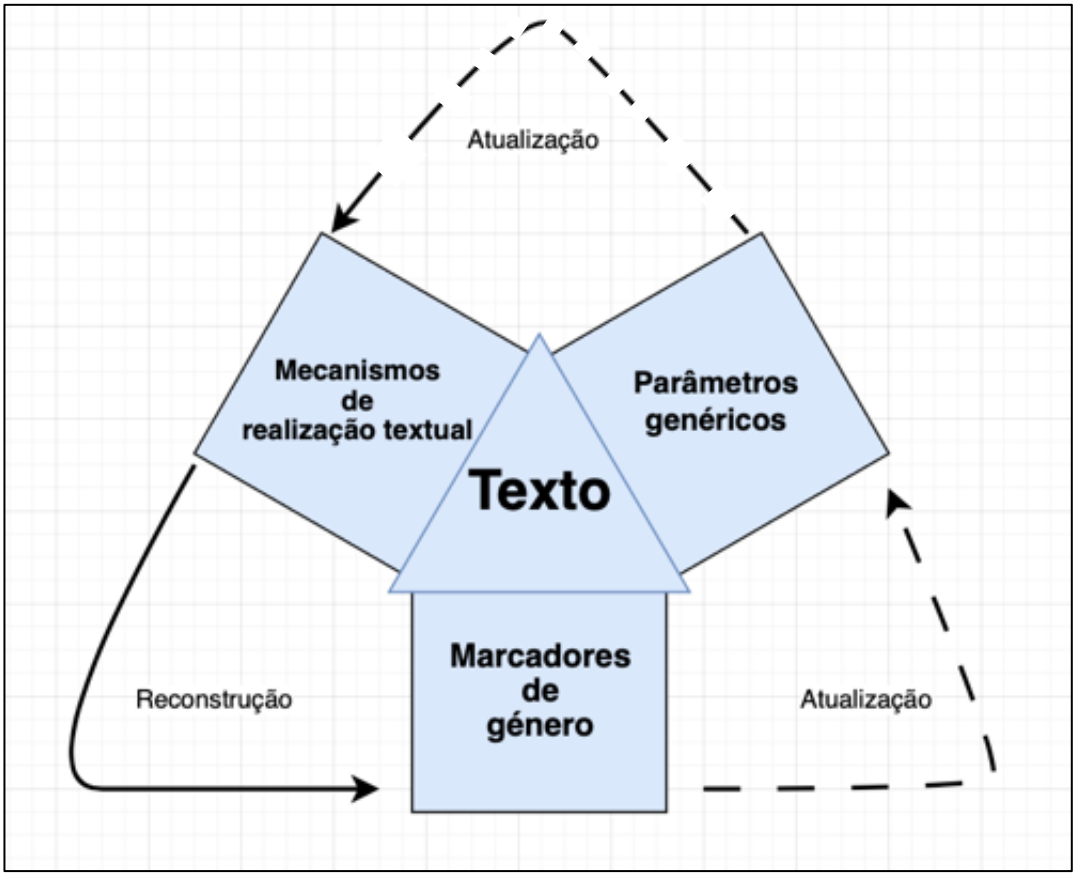


Figura 2: Relação entre mecanismos de realização textual, parâmetros genéricos e marcadores de género. Adaptado de Miranda (2010: 155)

De acordo com Coutinho & Miranda (2009: 42), podemos identificar dois grupos de marcadores de género que se distinguem pela forma explícita ou implícita em que inscrevem os textos em determinado género: os marcadores

autorreferenciais que exprimem explicitamente a categoria genérica do texto, e os marcadores inferenciais que, inscrevendo implicitamente os textos numa dada categoria genérica, necessitam de um trabalho de interpretação que ative os conhecimentos do interpretante. Os marcadores autoreferenciais surgem, embora não exclusivamente, sob a forma de etiquetas peritextuais ou integradas no corpo textual, e que geralmente se revelam suficientes para enquadrar os textos em termos genéricos, com algumas exceções (Miranda, 2007). Os marcadores inferenciais ocorrem de modo mais diverso, em unidades, mecanismos ou processos, que operam a vários níveis semiolinguísticos:

> *(…) os marcadores inferenciais são indícios que o recetor deve apreender, na maioria dos casos, de forma interligada. Ou seja, não funcionam isoladamente.* (Miranda, 2010: 156)

No entanto, estes níveis de análise podem ser organizados em três grupos de marcadores (Coutinho, 2019; Miranda, 2010): (i) os marcadores enunciativos, que se referem à ancoragem situacional do texto, a responsabilização enunciativa e a organização espácio-temporal, (ii) os marcadores temáticos que dizem respeito ao tema e à progressão temática e (iii) os marcadores composicionais que se referem aos fenómenos que permitem estruturar as unidades textuais.

Os marcadores que acabámos de referir operam ao nível da estrutura de cada texto. Na secção seguinte vamos explicitar como é que estes marcadores se organizam no texto de acordo com o modelo da arquitetura interna dos textos proposto pelo ISD.

1.2.4. A arquitetura interna dos textos

Bronckart (1997) propõe, além do modelo da ação da linguagem, um modelo para a análise da arquitetura interna dos textos. O modelo proposto concebe a organização do texto como um "folhado", com três camadas sobrepostas, e que permite mostrar uma hierarquia que releva mais de uma necessidade metodológica do que de uma distinção real, e que se relacionam diretamente com o grau de dependência contextual, permitindo, por isso, falar de camadas mais ou menos profundas.

Nível	Conteúdo	Função
Mecanismos de Responsabilização Enunciativa	• Gestão de vozes • Modalizações	Coesão Pragmática
Mecanismos de Textualização	• Coesão nominal • Coesão verbal • Conexão	Coesão Temática
Infraestrutura Geral	• Plano de texto • Sequências • Tipos discursivos	Planificação

Figura 3: Arquitetura interna dos textos, a partir de Bronckart (1997: 120).

O primeiro nível diz respeito aos mecanismos de responsabilização enunciativa, e funciona de modo quase independente, uma vez que estes se situam no nível mais superficial. A gestão das vozes, de acordo com Bronckart (1997: 120), remete para os mecanismos que estão diretamente ligados à interação entre o sujeito produtor e o(s) destinatário(s), e dividem-se em instâncias supraordenadas (narrador ou expositor) e instâncias infraordenadas (voz do autor, vozes sociais e vozes de personagens). As modalizações são mecanismos de caráter avaliativo ou judicativo sobre o conteúdo temático do texto e ocorrem a partir das vozes.

Vejamos os exemplos de quatro textos recolhidos do nosso *corpus* e disponível no Anexo 2. Na primeira tabela encontramos os aspetos situacionais, correspondentes ao contexto em que são produzidos os textos. Dada a natureza no nosso *corpus* (Cf. a secção **4.1.**), não existe grande variabilidade no contexto de

produção, uma vez que os textos foram recolhidos dentro da mesma atividade. Os produtores do texto estão identificados, assim como a temporalidade (assumindo o momento em que foi publicado). Sendo textos que circulam na comunicação social escrita, são desconhecidos os locais físicos de produção e os recetores. No que diz respeito aos parâmetros sociossubjetivos, observamos que também não existe grande variação: os enunciadores são jornalistas especializados em determinadas temáticas (cultura, tecnologia e política) e têm como destinatários tanto um público geral como um público mais especializado. Todos os textos têm como finalidade apresentar um acontecimento e/ou um produto, e o quadro social de circulação varia entre o jornal diário nacional, a revista mensal de especialidade e o sítio de informação nacional.

Na segunda camada situam-se os mecanismos de textualização que são responsáveis pela marcação, a estruturação e a progressão do conteúdo temático. Os mecanismos de textualização dividem-se ainda em mecanismos de coesão nominal, responsáveis pela introdução / recuperação de temas e personagens, mecanismos de coesão verbal, que têm como função a organização temporal dos acontecimentos, e os mecanismos de conexão, que articulam a progressão temática através dos organizadores textuais (cf. Coutinho (2019: 56).

Recuperando os nossos textos exemplares, é neste nível que começamos a observar algumas diferenças. No **Texto** 1, o tema do comentário é a musealização da arquitetura em Portugal e o anúncio da abertura da Casa da Arquitetura. Sendo o tema a arquitetura, o léxico utilizado é abundante em nomes próprios, tanto toponímicos como onomásticos. A coesão verbal e aspetual é assegurada através da presença de formas verbais no Modo Conjuntivo, introduzidas por construções impessoais que expressam conselhos e opiniões, enquanto o Modo Indicativo é usado para descrever o tempo presente (Presente e Pretérito Perfeito Composto). O **Texto 2** tem como tema a apresentação de um evento internacional de tecnologia (CES 2015) e, por isso, o seu léxico é mais técnico e especializado em eletrónica de consumo. A coesão verbal e aspetual é organizada com recurso predominante ao Pretérito Perfeito Simples do Indicativo e ao Pretérito Mais-Que-Perfeito do Indicativo. O Presente do Indicativo assume, neste texto, um valor gnómico. O **Texto**

4 é um comentário político (assinalado com uma etiqueta peritextual) e que tem um vocabulário predominantemente associado à política. Relativamente à coesão temporal e aspetual, esta é quase exclusivamente assegurada pelo tempo Presente do Indicativo com um valor deítico[7].

Aquilo que podemos observar nestes curtos exemplos é que existe mais variação quanto mais profundo é o nível de análise. Enquanto no primeiro nível as diferenças são muito ténues (ou nenhumas) porque têm contextos de produção semelhantes, no segundo nível ocorrem algumas diferenças, sobretudo ao nível da coesão temporal e aspetual, mas também do tipo de léxico selecionado. O **Texto 1** é indiscutivelmente um texto sobre arquitetura mas o seu léxico não é aquele que seria esperado num texto sobre arquitetura, como aquele que é usado nos **Textos 2**, **3** e **4**.

No **Texto 1** incluem-se expressões multipalavra como "arquitetura nacional", "museu de arquitetura" e "escola de arquitetura", assim como nomes próprios que remetem para o universo da arquitetura como "Prémio Pritzker", "Siza", "Souto de Moura" ou ainda instituições como "Casa da Arquitetura". O texto centra-se na menção de figuras e prémios, instituições e eventos associados diretamente à arquitetura, mas não existe vocabulário técnico específico de arquitetura, como nos outros textos.

A última camada situa-se no nivel mais profundo e, por isso, é a menos dependente do contexto da ação. A infraestrutura geral é constituída pelo plano geral do texto, que diz respeito à organização do conteúdo temático e que, de acordo com Bronckart (1997: 121), é recuperado no processo de leitura, podendo ser codificado num resumo. Desta última camada fazem também parte as sequências e os tipos discursivos. As sequências referem-se a "modos de planificações" que ocorrem dentro do plano geral do texto (sequências narrativas, explicativas, entre outras) e que seguiam o modelo de Adam (1992), mas que Bronckart (2008) acaba por

[7] O Presente do Indicativo ocorre sobretudo num contexto de citação e que, por isso, remete para o momento da enunciação.

abandonar por vários motivos[8]. O último elemento da infraestrutura geral são os tipos discursivos. Sendo um dos elementos mais desenvolvido no ISD, e também um dos principais contributos para o modelo de análise, necessitam de uma explicação mais detalhada, uma vez que são fundamentais para a análise que vamos empreender.

1.2.5. Tipos Discursivos

Como vimos anteriormente, Bronckart (1997, 2008) define o texto como uma unidade de produção verbal situada, com um fim e autossuficiente em termos de ação ou comunicação. Essa unidade pode ser classificada dentro de um género de texto, que por sua vez é uma referência organizacional para o texto. Os textos, por sua vez, são constituídos por segmentos, formados por formas linguísticas específicas, que se articulam na sua constituição. Estes segmentos resultam de um trabalho de semiotização (ou de construção discursiva) e são chamados de discurso. Os discursos, por sua vez, são identificados pelas regularidades linguísticas especificas e classificados em tipos discursivos, sendo parte integrante dos textos e dos géneros.

Os TD resultam, portanto, da interface entre as operações psicológicas que constroem os mundos discursivos e os fenómenos linguísticos que constituem a face visível desse mundo: as operações psicológicas são arquétipos genéricos e universais, enquanto os fenómenos linguísticos estão dependentes das regras das línguas naturais.

> *L'expression de type linguistique désigne le type de discours tel qu'il est effectivement sémiotisé dans le cadre d'une langue naturelle, avec ses propriétés morphosyntaxiques et sémantiques particulières. L'expression d'archétype psycologique désigne quant à elle cette entité abstraite ou ce construct, qu'est le type de discours appréhendé sous le seul angle des opérations psychologiques "pures" (...)* (Bronckart; 1997: 158)

[8] Entre os motivos apontados por Bronckart (2008: 52-53) destacamos o alcance limitado das sequências quando aplicadas a textos mais extensos, e o facto de esta abordagem composicional implicar uma análise do tipo *bottom-up*, ou seja, da unidade micro à unidade macro.

Os tipos discursivos englobam tanto os arquétipos psicológicos (mundos virtuais diferentes do mundo real do sujeito produtor) como os tipos linguísticos, sendo estes últimos a face visível dos primeiros e o modo como podemos aceder-lhes. Os tipos discursivos emergem na interação entre as coordenadas que estruturam o conteúdo temático presente no texto e as coordenadas do mundo quotidiano[9], relacionadas com a situação de enunciação. Dos diferentes planos de enunciação emerge uma correlação entre a organização temporal (conjunção ou disjunção da temporalidade construída no texto relativamente ao ato enunciativo) e uma relação com o agente (implicação ou autonomia do produtor textual relativamente ao ato enunciativo). O cruzamento dos dois eixos "constrói" quatro mundos discursivos (os arquétipos psicológicos), Expor implicado, Expor autónomo, Narrar implicado, Narrar autónomo, cada um com as suas marcas linguísticas específicas, e dependentes das contingências da língua natural em causa (os tipos linguísticos); esses mundos discursivos possuem um correspondente linguístico, os tipos discursivos - Discurso interativo, Discurso teórico, Relato interativo e Narração, sintetizados na Tabela 1.

		Organização temporal	
		Conjunção	Disjunção
		EXPOR	NARRAR
Organização atorial	Implicação	**Discurso interativo (DI)**	**Relato Interativo (RI)**
	Autonomia	**Discurso Teórico (DT)**	**Narração (N)**

Tabela 1: adaptado de Bronckart (1997: 159)

A organização temporal refere-se à localização dos acontecimentos relativamente à situação de produção. Assim, se os conteúdos temáticos descrevem acontecimentos simultâneos ao momento de enunciação, ocorre um valor de

[9] Em francês "*monde ordinaire*".

conjunção temporal. Mas se, contrariamente, os acontecimentos não são simultâneos ao momento de enunciação, ocorre um valor de disjunção temporal. A conjunção temporal produz dois tipos discursivos da ordem do *Expor*: o Discurso Interativo (DI) e o Discurso Teórico (DT), enquanto a disjunção temporal produz dois tipos discursivos da ordem do *Narrar*: o Relato Interativo (RI) e a Narração (N). Para além da organização temporal, importa observar a relação de agentividade que está presente nos textos e que assume duas vertentes: a implicação e a autonomia. No primeiro caso, as instâncias agentivas do texto inscrevem-se nos parâmetros da ação de linguagem do texto, e temos os tipos discursivos DI e RI. Quando as instâncias agentivas não se inscrevem nos parâmetros da ação de linguagem, obtemos os tipos discursivos DT e N. Há, assim, dois tipos de relação temporal e agentiva cujo cruzamento dá, por sua vez, origem a quatro possibilidades que se encontram sintetizadas na Tabela 2:

<table>
<tr><td colspan="2">Relação entre as coordenadas que organizam o conteúdo temático de um texto e as coordenadas do mundo normal:</td></tr>
<tr><td>Relação de disjunção</td><td>Relação de conjunção</td></tr>
<tr><td>a organização das representações mobilizadas, depende de uma origem espacio-temporal que especifica o tipo de disjunção operada; trata-se, neste caso, de narrar os factos.</td><td>na ausência de uma origem explícita, as representações mobilizadas organizam-se em função das coordenadas do mundo em que decorre a ação (ou mundo normal); trata-se neste caso de mostrar ou expor os factos.</td></tr>
<tr><td colspan="2">Relação entre as instâncias de agentividade (personagens, grupos, instituições, etc.) com a respetiva inscrição espacio-temporal e os parâmetros físicos da ação (agente produtor, interlocutor, espaço e tempo de produção)</td></tr>
<tr><td>Relação de implicação</td><td>Relação de autonomia</td></tr>
<tr><td>se, através de ocorrências deíticas, o texto mobiliza ou implica os parâmetros físicos (necessitando assim a interpretação de ter acesso às condições de produção);</td><td>se, na ausência de qualquer explicitação, as instâncias de agentividade permanecem independentes dos parâmetros físicos da ação (não sendo necessário, para a</td></tr>
</table>

	interpretação, o conhecimento das condições de produção)

Tabela 2: Extraído de Coutinho (2019: 48)

Como já referimos anteriormente, é através dos tipos linguisticos que constituem cada um destes mundos discursivos, e que estão dependentes das línguas naturais, que podemos aceder aos TD. A identificação e caracterização dos TD para a língua francesa contemporânea foram feitas a partir da análise quantitativa e qualitativa de textos empíricos (Bronckart, 1997: 80-86), com recurso a um *corpus* de textos escritos e orais. Na secção seguinte vamos dar conta dos tipos linguísticos que estão associados a cada um dos tipos discursivos.

1.2.6. Unidades linguísticas: propriedades linguísticas dos tipos discursivos

Como referimos anteriormente, os tipos discursivos emergem da correlação entre dois eixos que correspondem a planos diferentes de enunciação. Na Tabela 3 podemos observar os tipos linguísticos que ocorrem em cada um dos tipos discursivos.

Expor Implicado	Discurso Interativo
⇒ Dialogado ou monologado, oral ou escrito. ⇒ Turnos de fala nas formas dialogadas. ⇒ Presença de unidades que remetem para a interação verbal (real ou encenada). ⇒ Presença de frases não declarativas (interrogativas, exclamativas e imperativas). ⇒ Verbos do plano do discurso (Benveniste): presente, pretérito perfeito e futuro perifrástico; geralmente, **com valor deíctico**. ⇒ Presença de unidades que remetem para objectos acessíveis no espaço (**deícticos espaciais**) e no tempo (**deícticos temporais**). ⇒ Nomes próprios, verbos e pronomes de 1ª e de 2ª pessoa, do singular e plural, que remetem para os protagonistas da interação verbal (valor exofórico).	

Narrar Implicado	Relato Interativo
⇒ Geralmente monologado. ⇒ Ausência de frases não declarativas. ⇒ Exploração do subsistema de verbos do plano da história (Benveniste) ou dos tempos narrativos (Weinrich): pretérito perfeito, imperfeito, mais-que-perfeito, futuro simples e condicional. ⇒ Presença de organizadores temporais (advérbios, sintagmas preposicionais, coordenativos, subordinativos, etc.). ⇒ Presença de pronomes de 1ª e de 2ª pessoa do singular e do plural, que remetem para os protagonistas da interação verbal. ⇒ Presença dominante de anáforas pronominais, às vezes associadas a anáforas nominais (repetição fiel do antecedente). ⇒ Densidade verbal elevada. ⇒ Densidade sintagmática baixa.	

⇒ Presença do pronome indefinido "on" [se], com valor de primeira pessoa do singular ou do plural. ⇒ Presença de anáforas pronominais. ⇒ Presença de auxiliares de modo (poder, dever, querer, ser preciso, etc.). ⇒ Densidade verbal elevada. ⇒ Densidade sintagmática baixa.	

Expor Autónomo	Discurso Teórico
⇒ Monologado e escrito. ⇒ Ausência de frases não declarativas. ⇒ Exploração do subsistema de verbos do plano do discurso (Benveniste), mas com uma clara dominância das formas do presente e do pretérito perfeito com valor genérico. ⇒ **Ausência de unidades que remetam aos interactantes ou ao espaço-tempo da produção**. ⇒ Possibilidade de ocorrência da segunda pessoa do plural ou da forma "on", quando não remetem para os participantes da interacção em curso. ⇒ Organizadores com valor lógico e argumentativo. ⇒ Modalizações lógicas e do auxiliar "poder". ⇒ Procedimentos de focalização (metatextuais, intratextuais, intertextuais). ⇒ Presença de frases passivas. ⇒ Presença de anáforas pronominais, anáforas nominais e procedimentos de referenciação deíctica intratextual. ⇒ Densidade verbal baixa. ⇒ Densidade sintagmática elevada.	

Narrar Autónomo	Narração
⇒ Geralmente escrito e sempre monologado. ⇒ Presença exclusiva de frases declarativas. ⇒ Exploração do subsistema de verbos do plano da história (Benveniste) ou dos tempos narrativos (Weinrich), sendo o pretérito perfeito e o imperfeito os tempos dominantes. ⇒ Presença de organizadores temporais (advérbios, sintagmas preposicionais, coordenativos, subordinativos, etc.). ⇒ **Ausência de pronomes de 1ª e de 2ª pessoa do singular e do plural, que remetem para os protagonistas da interação verbal**. ⇒ Presença conjunta de anáforas pronominais e anáforas nominais (geralmente, retomada do sintagma antecedente com substituição lexical). ⇒ Densidade verbal média. ⇒ Densidade sintagmática média.	

Tabela 1: Extraído de Miranda (2010: 139)

As configurações linguísticas que se apresentam na Tabela 3 mostram também que um determinado tipo linguístico pode ocorrer em mais do que um tipo discursivo, o que significa que não é suficiente, para identificar o tipo discursivo, analisar a presença ou ausência de um único tipo linguístico, mas observar como os vários tipos linguísticos presentes no texto interagem.

Bronckart (1997) considera que os textos raramente são constituídos por um único TD mas por vários tipos discursivos, em que um é predominante sobre os outros, articulando-se entre si através de processos que podem variar entre o encaixe e a fusão:

> *De tels textes comportent quasi nécessairement un type majeur et un ou plusieurs types mineurs ou subordonnés (d'autres types pouvant d'ailleurs être subordonnés à ces types mineurs, en un processus d'emboîtement potentiellement infini). Et l'articulation d'un type mineur au type majeur peut s'effectuer selon deux modalités générales: l'enchâssement et la fusion.* (Bronckart, 1997)

1.2.7. Articulação dos tipos discursivos

Um texto é considerado homogéneo se constituído de um único tipo discursivo (o género entrada de dicionário, por exemplo, é constituido pelo Discurso Teórico), enquanto um texto heterogéneo é constituido por mais do que um TD. De acordo com Bronckart (1997: 258), os textos homogéneos são pouco frequentes ou mesmo raros, podendo estabelecer-se uma relação direta entre o tamanho do texto e o número de TD presentes: a probabilidade de ocorrerem vários TD num texto aumenta quanto maior for o texto. A relação entre os TD nem sempre é igual uma vez que é frequente ocorrer uma hierarquização dos TD, em que um TD surge como dominante e os restantes surgem como subordinados. A articulação entre os diferentes TD é um dos elementos que confere a coesão textual, e pode ocorrer de dois modos:

(i) Encaixe: os tipos discursivos surgem delimitados por marcas (gráficas, lexicais ou morfosintáticas) que hierarquizam claramente os TD em dominante e subordinados.

(ii) Fusão: os tipos discursivos integram-se sem marcas claras entre eles e sem demonstrar uma hierarquia.

Embora tenhamos mostrado na Tabela 2 que os tipos discursivos se organizam em torno de dois eixos (atorial e temporal), colocando o peso dessa distinção numa decisão binária, a verdade é que por vezes ocorrem sobreposições que se traduzem na articulação em fusão (Bronckart, 2008: 73). Vejamos um exemplo:

> *Há algum tempo, o alto patrocinador desta conferência disse-nos que via mais vida para além do orçamento. E muitos de nós pensaram: "Graças a Deus". Serviu este painel para nos recordar que, do mesmo modo, há mais língua portuguesa para*

além da falada em Portugal. (...) Em Lisboa, como em São Paulo e em Maputo, é uma língua que veio de fora. O que não impede os portugueses de afirmar que a língua lhes pertence; impede-os, sim, de afirmar que a língua só a eles pertence como ainda se ouve dizer. (Texto 795)

Este excerto de uma conferência, escrita e publicada como um comentário no âmbito académico, começa com um conjunto de elementos que se situam próximos do Discurso Interativo, nomeadamente a locução temporal "há algum tempo" (que remete temporalmente para o momento da enunciação do discurso oral), os pronomes pessoais de primeira pessoa "nos" e "nós", que reenviam para o autor e destinatário do texto, entre outros. No entanto, também se observam elementos linguísticos próprios da Relato Interativo, como os verbos no Pretérito Perfeito e no Pretérito Imperfeito do Indicativo. O texto continua com segmentos do Discurso Teórico pontuados por elementos do Discurso Interativo e do Relato Interativo, numa amálgama de TD que se sobrepoem, e que têm como objetivo captar a atenção do destinatário.

1.2.8. Relação entre tipos discursivos e genéro textual

A relação entre tipos discursivos e género textual é uma questão que foi abordada por Miranda (2008) que, para analisar esta relação, descreve três níveis de relação entre estes dois elementos, tentando descrever como interagem e se condicionam mutuamente na produção textual. Os três níveis de relação são os seguintes:

(i) O primeiro nível descreve a relação constitutiva que existe entre tipos discursivos e género textual: é uma relação expressa na própria definição de tipo discursivo feita por Bronckart (1997: 139): *Nous avons em conséquence soutenue que c'étaient ces segments constitutifs d'un genre qui devaient êtres considérés comme des types linguistiques*. Os tipos discursivos não são elementos opcionais ou secundários em relação aos géneros textuais porque a sua presença é intrínseca (*constitutifs*) a qualquer género textual. Deste modo, cada género textual, ao ser produzido, está necessariamente ligado a um ou mais tipos discursivos, que são essenciais para a sua configuração.

(ii) O segundo nível concerne a escolha e estabilização dos tipos discursivos: este nível releva da importância que os géneros textuais têm na seleção dos tipos discursivos

que os constituem, e como essa escolha se reflete numa certa estabilização dessa escolha. Trata-se, portanto, de um nível que permite articular a transversalidade dos tipos discursivos (o mesmo tipo discursivo ocorre em vários géneros textuais) com a relativa estabilidade que os géneros proporcionam: *Neste sentido, os tipos de discurso (e as suas modalidades de articulação) mostram alguma regularidade relativamente à sua ocorrência em diversos géneros e até podem funcionar como pistas para a identificação dos géneros* (Miranda, 2008: 89).

(iii) O terceiro nível de relação incide sobre a relação entre géneros textuais e tipos linguísticos, e é o nível mais desenvolvido. A autora retoma a noção de tipo discursivo, dividida em duas vertentes: uma psicocognitiva, relacionada com mundos discursivos ou "atitudes de locução", e uma semiótica, associada às unidades linguísticas que traduzem as operações psicocognitivas. Partindo destas duas vertentes, a questão central que se coloca é se os géneros, ao selecionarem o(s) tipo(s) discursivo(s) que os constituem, instanciam as unidades linguísticas próprias desse tipo discursivo de maneira idêntica. Os exemplos estudados mostram que as unidades linguísticas mobilizadas podem ter valores diferentes, dependendo do género em que ocorrem.

Em Miranda (2010) é, igualmente, recuperada a questão da relação entre tipos discursivos e género, perspetivando que, além de assumir o tipo discursivo como parte integrante dos géneros textuais, também o plano textual tem uma papel central na organização dos tipos discursivos.

1.2.9. Tipos discursivos e organização textual

Como já referimos anteriormente, a infraestrutura geral dos textos corresponde ao nível menos dependente do contexto da ação, e é composta pelo plano do texto, pelos tipos discursivos e pelas sequências e outras formas de planificação. Já referimos, também, que os tipos discursivos se articulam entre si pela fusão ou encaixe e que o plano geral dos textos surge definido como a organização temática dos conteúdos, e que é determinado pela combinação dos tipos discursivos, entre outras formas de planeamento:

> *Enfin et surtout parce que ce plan général est surdéterminé par la combinatoire spécifique des types de discours, des séquences et des autres formes de planification apparaissant dans le texte*. (Bronckart, 1997: 253)

Mas existe uma dificuldade em tipificar o plano geral dos textos, assumida pelo autor, e que surge pela forma como os vários fatores que contribuem para este plano se cruzam. Convém dizermos que, até este momento, estamos perante uma planificação que se organiza verticalmente, entre o género e os tipos discursivos convocados.

A organização horizontal dos tipos discursivos organiza-se, de acordo com Bronckart (2008: 80-81), através dos mecanismos de conexão e coesão nominal que desempenham um papel fundamental na manutenção da coerência temática do texto, e que não estão confinados às fronteiras dos tipos discursivos. A conexão tem como função principal tornar visível a articulação dentro do texto, como as que ocorrem entre os tipos de discurso, e a coesão nominal evidencia os elos entre "argumentos" (em oposição aos predicados). Do ponto de vista do significado, a conexão pode ser subdividida em operações como segmentação, diferenciação e ligação, dependendo do tipo de articulação envolvida. A coesão nominal, por sua vez, pode ser vista nas operações de introdução de novos argumentos ou de retoma de argumentos anteriores. Em termos do significante, os mecanismos de conexão e coesão são marcados por unidades linguísticas específicas, que são usadas de forma preferencial para indicar suboperações específicas, embora nem sempre essa correspondência seja unívoca, já que algumas marcas podem representar diferentes suboperações simultaneamente.

No entanto, a questão que nos colocamos é se estas operações de conexão e coesão nominal conduzem a outro tipo de articulação entre os tipos discursivos, em que os tipos discursivos dominantes bloqueiam a ocorrência de outros tipos discursivos específicos, ou se existe uma correlação (positiva ou negativa[10]) entre os tipos discursivos.

[10] Adotamos os termos estatísticos de "correlação positiva" e "correlação negativa": "Dizemos que duas variáveis, X e Y, estão positivamente correlacionadas quando elas caminham num mesmo sentido (...). Estão negativamente correlacionadas quando elas caminham em sentidos opostos, (...) (Barbetta, 2002: 271).

1.3. A *deixis*

Como referimos no capítulo anterior, os tipos discursivos englobam tanto os arquétipos psicológicos como os tipos linguísticos e organizam-se em torno de dois eixos, temporal e agentivo, que constroem os quatro tipos discursivos. Referimos, também, que só podemos aceder aos arquétipos psicológicos através dos tipos linguísticos, sendo necessário identificar quais destes tipos linguísticos funcionam nas operações de referenciação espácio-temporal e enunciativa dos TD. Na tabela 3, recuperámos a síntese feita por Miranda (2010) com os tipos linguísticos que ocorrem em cada um dos tipos discursivos, e podemos ver que a presença / ausência de deíticos podem auxiliar na distinção entre os quatro tipos discursivos, uma vez que estes estabelecem coordenadas com o mundo real. Assim, no que segue, vamos atender a *deixis* (pessoal, espacial e temporal) de modo a elencar as pistas linguísticas pertinentes para a deteção dos tipos linguísticos.

1.3.1. A *deixis* - Pessoa

No contexto deste trabalho, utilizaremos a definição proposta por Lyons (1977: 637) para a *deixis*:

> *By deixis is meant the location and identification of persons, objects, events, processes and activities being talked about, or referred to, in relation to the spatiotemporal context created and sustained by the act of utterance and the participation in it, typically, of a single speaker and at least one addressee.*

A *deixis* é o processo através do qual é feita a construção de valores linguísticos, que resultam na representação da referência, relativamente aos sujeitos e ao espaço-tempo no enunciado e no texto. Este processo é materializado linguisticamente através dos deíticos, enquanto "gestos verbais cuja função primária é estabelecer a ligação entre o explícito e o implícito na comunicação verbal". (Fonseca, 1992: 70). Não estando no âmbito deste trabalho fazer um estudo aprofundado sobre a *deixis*[11], parece-nos, no entanto, relevante abordar alguns aspetos para mostrar a sua complexidade, e para podermos estabelecer algumas ferramentas de análise.

[11] Cf. (Fonseca, 1992; Levinson, 2006; Lyons, 1977).

No capítulo referente à *deixis*, Lyons (1977: 638-639) começa por destacar a ambiguidade e a indeterminação dos enunciados que, sendo facilmente interpretáveis em situações de enunciação canónicas, podem ser sujeitos a várias formas de ambiguidade ou indeterminação quando produzidos em situações não canónicas, como a separação espacial e/ou temporal dos participantes ou se o enunciado é escrito ou oral. A ambiguidade que releva das situações não canónicas decorre, como refere o autor, de um egocentrismo enunciativo em que o ponto-zero espaciotemporal é determinado pelo lugar do enunciador, no momento do enunciado. O termo egocentrismo é introduzido para ilustrar o facto de o falante se colocar no papel de *ego* e relacionar tudo a partir de seu ponto de vista. O egocentrismo é tanto temporal quanto espacial, e a mudança de papel do falante afeta a localização espaciotemporal da referência (aqui e agora) e influencia não só a marcação do tempo, mas também a categoria gramatical de pessoa. O autor discute como os papéis dos participantes numa situação de comunicação são gramaticalizados nas línguas, o que resulta na distinção entre as diferentes formas de pronomes pessoais, como "primeira pessoa", "segunda pessoa" e "terceira pessoa". No entanto, Lyons [12]considera que só a primeira e segunda pessoa possuem valores deíticos:

> *It is important to note, however, that only the speaker and addressee are actually participating in the drama. The term 'third person ' is negatively defined with respect to 'first person' and' second person': it does not correlate with any positive participant role. (Lyons, 1977: 638)*

Levinson (2006) expande o conceito de *deixis* para incluir também os deíticos sociais, que incluem, por exemplo, as formas de tratamento. No caso da primeira pessoa, ocorre quando há uma coincidência entre o sujeito enunciador e o sujeito sintático, enquanto na segunda pessoa, ocorre quando há divergência entre o sujeito enunciador e o sujeito sintático que coincide com o coenunciador. Importa notar que o autor procura mostrar, também, como a noção de pessoa é refletida e manifestada nas estruturas gramaticais das línguas naturais, influenciando a forma como os falantes se referem a si mesmos, ao interlocutor e a outras entidades mencionadas na comunicação.

[12] Cf. também Valentim (2015).

		Sujeito	Objeto Direto	Objeto Indireto	Complemento Oblíquo	Agente da passiva
Singular	1ª pessoa	*eu*	*me*	*me*	*mim, comigo*	*mim*
	2ª pessoa	*tu*	*te*	*te*	*ti, contigo*	*ti*
Plural	1ª pessoa	*nós*	*nos*	*nos*	*nós, connosco*	*nós*
	2ª pessoa	*vós / vocês*	*vos*	*vos*	*vós, convosco*	*vós*

Tabela 2: Extraídos de Valentim (2015: 300)

Na Tabela 4 podemos observar o paradigma dos pronomes pessoais em Português Europeu (PE) que podem ter um valor deítico. Como a autora sublinha (Valentim, 2015), em PE os pronomes pessoais apresentam flexão em número, género e pessoa, além de poderem ser marcados morfologicamente através da flexão verbal, não sendo obrigatória a sua presença na frase.

> **(i)** *Chegámos à CES 2015, Consumer Electronics Show, de Las Vegas, com alguma expectativa sobre os televisores quantum do LED*. (Texto 2)
>
> **(ii)** *O entusiasmo esmoreceu ao confirmarmos que o dito novo tipo de iluminação de ecrã já tinha sido usado nalguns modelos da série Triluminos da Sony, em 2013*. (Texto 2)
>
> **(iii)** *Tínhamos testado alguns e confrontámos a teoria com os resultados obtidos na altura, em laboratório*. (Texto 2)

Estes exemplos retirados do Texto 2 contido no Anexo 1, mostram como o sujeito das frases é expresso através da flexão verbal. As marcas de primeira pessoa do plural fazem coincidir o falante (o produtor físico do comentário) com o sujeito do enunciado, sobre quem recai a ação. Como vimos anteriormente, os tipos discursivos organizam-se em torno de dois eixos, sendo um deles a relação com o agente - implicação ou autonomia do produtor textual relativamente ao ato enunciativo - que, nos exemplos, é de implicação, situando esta sequência textual como Discurso Interativo (DI) ou Relato Interativo (RI). Na secção seguinte vamos analisar o segundo eixo, que diz respeito à organização temporal do enunciado.

1.3.2. A *deixis* - Tempo

Para Lyons (1977), o conceito de tempo verbal não é uma categoria exclusiva da flexão do verbo: o autor argumenta que semanticamente, o tempo é uma categoria que se aplica à frase como um todo, e às orações que estejam relacionadas com a categoria tempo dentro de uma frase.

> *(...) the participants in a language-event must be able to control and interrelate at least two different frames of temporal reference: the deictic and the non-deictic*. (Lyons, 1977: 678)

Significa, portanto, que os falantes de uma língua são capazes de lidar com dois quadros temporais distintos: o quadro de referência deítico, que se refere ao tempo em relação à perspetiva do falante no momento da enunciação, e o quadro de referência não-deítico, que é independente da perspetiva do falante. O tempo verbal é uma parte do sistema de referência deítico, e a sua função é estabelecer a relação temporal entre o tempo da situação que está a ser descrita e o ponto de referência temporal zero dentro do contexto da enunciação. Embora o autor afirme que a categoria tempo não é universal, considera que todas as línguas têm deíticos de tempo para estabelecer uma distinção entre deítico e não-deítico. Para o PE concorrem na classe dos deíticos temporais advérbios como *agora*, *hoje*, *amanhã*, entre outras formas e expressões adverbiais que permitem localizar um evento no tempo, determinado em relação ao momento da enunciação. Observe-se o exemplo seguinte do Texto 1 do Anexo 1:

(i) *Agora que a Casa de Arquitectura abrirá a Norte no próximo Verão, com uma ambição internacional (...)*. (Texto 1)

Neste exemplo, a ancoragem temporal assenta numa localização deítica, com a presença de um advérbio de tempo ("Agora") e de uma expressão adverbial temporal ("no próximo Verão"), que só são interpretáveis relativamente ao momento de produção/publicação do texto (26/11/2016). A referência temporal obriga a um pré-conhecimento dos factos, localizados num espaço de tempo definido.

No português europeu (PE), a marcação de tempo não é feita apenas com recurso a advérbios e expressões de tempo, mas também através da flexão verbal. Referimos anteriormente que a flexão verbal em PE tem marcas de pessoa e de tempo,

entre outras informações. Não estando no âmbito deste trabalho fazer um estudo aprofundado sobre a temporalidade em PE, importa, no entanto, estabelecer alguns conceitos e distinções sobre este tópico. Vamos começar por distinguir entre tempo cronológico e tempo gramatical: usamos a expressão tempo cronológico para designar o valor de referência temporal do enunciado (eixo temporal relativamente ao momento de enunciação) e o tempo gramatical [13] para designar o valor temporal marcado pela flexão verbal (Valentim, 2015: 309) - anterioridade, simultaneidade e posteridade.

De acordo com Fonseca (1992) a atividade enunciativa gera um marco de referência temporal claro e inequívoco, e este marco de referência está centrado no presente, que é a coordenada do momento da enunciação. É através deste marco de referência temporal, centrado no sujeito falante e, por isso, deítico que é possível organizar os eventos num antes e/ou depois (passado ou futuro):

> *Presente, passado e futuro não existem senão como perspetivas assumidas por um sujeito falante a partir do momento em que está ou de um momento em que imagina estar*. (Fonseca, 1992: 176)

A organização deste presente, passado e futuro é feita a partir do momento da enunciação (T0), organizados linearmente. Vejamos um exemplo alargado do Texto 2:

> **(ii)** *Chegámos à CES 2015 , Consumer Electronics Show de Las Vegas , com alguma expectativa sobre os televisores quantum dot LED. O entusiasmo esmoreceu ao confirmarmos que o dito novo tipo de iluminação de ecrã já tinha sido usado em alguns modelos da série Triluminos da Sony, em 2013. Tínhamos testado alguns e confrontámos a teoria com os resultados obtidos na altura, em laboratório*. (Texto 2)

O primeiro elemento a considerar é o tempo verbal Pretérito Perfeito Simples do Indicativo ("chegámos", "esmoreceu" e "confrontámos") ao qual é atribuído um valor de anterioridade em relação ao momento da enunciação (T0). Concomitantemente, a presença de dois localizadores temporais autónomos (neste caso, as datas 2015 e 2013) irão determinar o ponto de partida das duas sequências temporais, a partir das quais se

[13] Tempo linguístico em Fonseca (1992: 175).

vai desdobrar a temporalidade semiotizada neste segmento textual: uma relativa ao eixo temporal de 2015 construída com o tempo verbal PPS e outra relativa ao eixo temporal de 2013 edificada com o Mais-que-Perfeito Composto do Indicativo ("tinha sido" e "tínhamos testado"). Existem, neste exemplo, relações de interdependência que constroem as referências temporais e que são atualizadas através de operações de referenciação, a saber uma referência temporal relativa ao momento de enunciação (T0), outra relativa ao localizador 2015 e uma terceira a 2013, retomada anaforicamente pela expressão temporal "na altura". O valor dos verbos (valores de anterioridade, posterioridade e simultaneidade) é atualizado em função destes localizadores e da relação que se estabelece entre eles. Como observa Valentim (2015: 309), embora o tempo gramatical (expresso pela flexão verbal) possa coincidir com o tempo cronológico nem sempre estas categorias convergem. O valor temporal dos enunciados é linear e inclui um ponto de referência (a situação de enunciação) em relação ao qual os eventos linguísticos são localizados. Este exemplo agora apresentado mostra como a interpretação temporal de cada uma destas situações, seja como eventos do passado, presente ou futuro, depende de um referencial absoluto, que é o momento em que o enunciador constrói esses eventos linguísticos.

1.3.4. A *deixis* - Espaço

A construção do sistema de referência deítico parte, como referimos anteriormente, do ponto-zero espaciotemporal determinado pelo lugar do enunciador, instituindo o contexto de enunciação. Para que este contexto se estabeleça entre o enunciador e o interlocutor, é necessário que haja uma convergência das coordenadas espaciotemporais entre ambos. Os deíticos espaciais são unidades linguísticas que fazem a ancoragem espacial, permitindo a interpretação do enunciado em relação ao seu espaço. A *deixis* espacial inclui não só os advérbios e expressões adverbiais de lugar, mas também os pronomes e determinantes demonstrativos que indicam proximidade ou afastamento do locutor e/ou recetor, e também verbos que demonstram movimento / localização *de* e *para* o espaço do emissor. Em PE, os pronomes e determinantes demonstrativos são construídos a partir da localização de dois referentes (eu, tu/você) ao qual estão associados três advérbios de lugar (aqui, aí e ali), e que podemos representar da seguinte maneira:

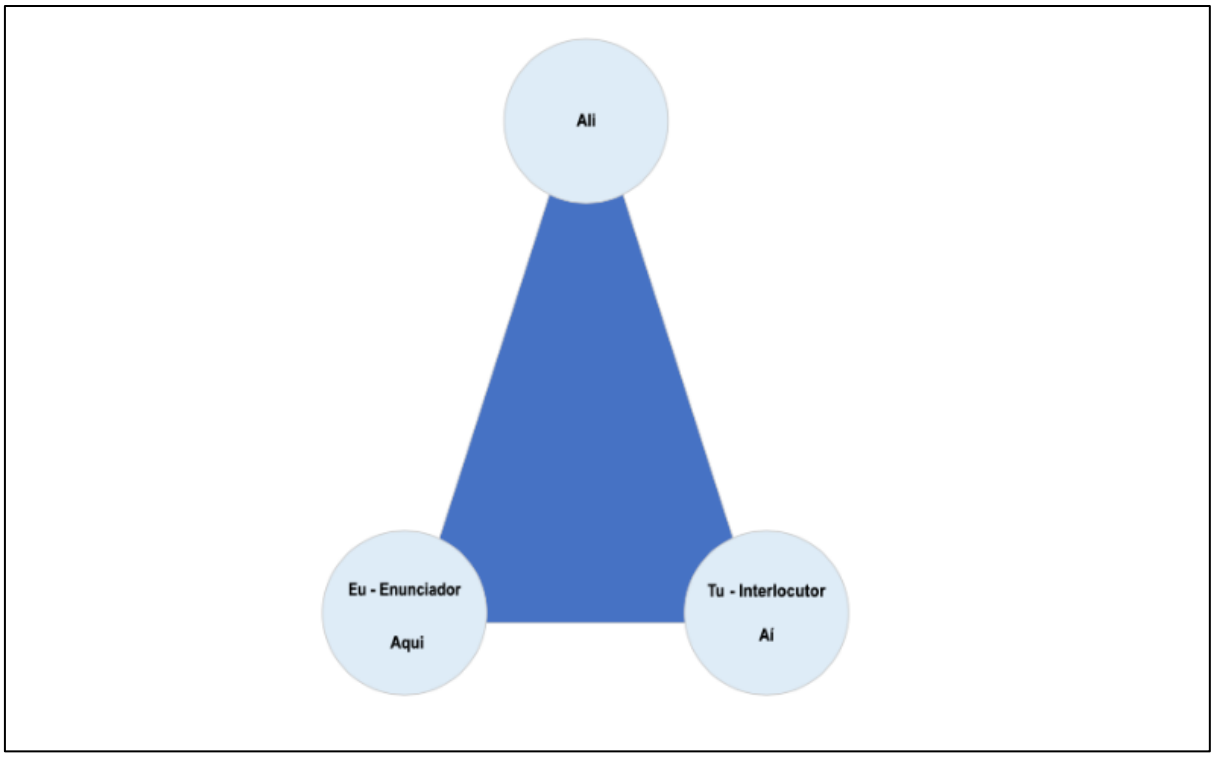


Figura 4: Representação gráfica da deixis espacial.

Na Figura 4 podemos observar que os deíticos espaciais, tal como os deíticos pessoais e temporais, são fixos a partir da referência egocêntrica que é o enunciador. Assim, o advérbio 'aqui' localiza o espaço junto ao enunciador, enquanto o advérbio 'aí' localiza o espaço junto do interlocutor, e, finalmente, o advérbio 'ali' localiza um espaço afastado do enunciador e do interlocutor.

Os demonstrativos, além das funções de determinante e pronome, podem ter uma função anafórica para identificar um construído linguístico prévio ou partilhado pelo conhecimento comum entre o enunciador e o interlocutor.

> (i) *Nuno Sampaio, que reivindica com a Casa de Arquitectura ter conseguido fazer em Matosinhos o primeiro museu só de arquitectura, lembra que* ***esta*** *é a única área relevante da Cultura em que o Estado não está presente (...).* (Texto 1)
>
> (ii) *Agora que a Casa de Arquitectura abrirá a Norte no próximo Verão, com uma ambição internacional, como mostra* ***esta*** *apresentação em Veneza (...).* (Texto1)

No exemplo (i) o demonstrativo ***esta*** faz a retoma do nome *arquitectura*, enquanto no exemplo (ii) ocupa a função de determinante da expressão nominal "apresentação em Veneza" que não é referida previamente, e não tem uma função locativa relativamente ao enunciado. É um acontecimento que faz parte do conhecimento partilhado do

enunciador e do interlocutor (leitores do artigo). O que nós observamos nos textos exemplares é que há um processo de transposição fictiva[14] do momento enunciativo, que permite ancorá-lo espacial e temporalmente num local distinto do aqui / agora. Este processo de transposição mental de um espaço e tempo distintos do momento da enunciação para um espaço imaginário convocado, permite que os deíticos possam ser utilizados para apontar uma situação ausente. Vejamos alguns exemplos:

> **(i)** *Escrutinámos os resultados dos modelos da Sony com* quantum dot *LED. Ao contrário dos televisores LCD LED, nos* quantum dot*, a perceção das cores divide um pouco o painel que visualiza as imagens dos televisores no nosso teste. Alguns dão boa nota às cores, mas outros relatam desvios nos tons. Sentados frente a um ecrã* quantum dot *para ver uma imagem só com as cores mais puras, ficaríamos impressionados com a tecnologia.* (Texto 2)
>
> **(ii)** *No habitual espaço de comentário, na SIC, o ex-presidente social-democrata apontou o caminho: o PSD tem de "ganhar as eleições autárquicas" e apostar em "grandes candidatos" para os principais centros urbanos. "Se ganhar as eleições autárquicas fica mais bem preparado para ganhar as eleições legislativas". Ou então pode haver uma "indesejável crise interna e de liderança", foi avisando o comentador. (Texto 4)*

No exemplo (i), há uma dissociação espaciotemporal entre o momento de enunciação (produção do texto) e o momento de receção / leitura do texto. No entanto, o jornalista evoca um espaço situacional distinto através de um localizador espacial autónomo (referido anteriormente no texto através da expressão "Chegámos à CES 2015 (...), de Las Vegas"), a partir do qual são construídas novas referências espaciotemporais: se por um lado há, temporalmente, um valor de anterioridade em relação ao momento da enunciação, por outro, a utilização do Presente do Indicativo (divide, dão, relatam...) sugere uma transposição para esse espaço/tempo que permite o uso de um presente com valor deítico, juntamente com a expressão "Sentados frente a" que *coloca* o leitor nesse espaço fictício. No exemplo (ii), retirado do Texto 4, encontramos uma situação diferente em que há o recurso a um localizador espacial "*No*

[14] A noção de transposição fictiva é aqui utilizado por nós, para designar o processo através do qual o sujeito falante desloca-se fictivamente para outras coordenadas espaciotemporais, criando os seus próprios referentes (Cf. Fonseca (1992: 135-140)).

habitual espaço de comentário, na SIC" que permite transpor a situação enunciativa para um espaço partilhado pelo conhecimento comum entre o enunciador e o interlocutor. É partir desse espaço e tempo partilhado que se desenvolve todo o enunciado, que recorre a processos grafológicos para delimitar um enunciado citado. A citação é um processo discursivo que assume, neste texto, duas funções: a introdução de um novo tópico de comentário e a responsabilização enunciativa que permite, por sua vez, imprimir um caráter objetivo ao texto. Neste exemplo, o uso das aspas permite articular dois tipos discursivos: o discurso teórico dominante e o discurso interativo intercalado, este último caracterizado pelo uso do Presente do Indicativo com valor deítico e de auxiliares de modo, juntamente com a presença das aspas que remetem para a interação verbal (real ou encenada).

Em resumo:

- A *deixis* é o processo de construção de valores linguísticos relacionados com a referência em relação a pessoas, espaço e tempo no discurso.
- A referenciação é materializada linguisticamente através de deíticos, que são gestos verbais usados para conectar o explícito e o implícito na comunicação verbal.
- A *deixis* parte do conceito de "egocentrismo enunciativo", onde o falante se coloca no papel de "ego" e relaciona a localização espacial e temporal a partir do seu ponto de vista.
- A transposição fictiva permite ancorar o enunciado, espacial e temporalmente, num local distinto do aqui / agora.

Como pudemos constatar nos exemplos retirados do *corpus*, os valores e funções que os tipos linguísticos assumem, bem como a interação com outros elementos textuais, dependem das relações de interdependência que constroem as referências enunciativas, temporais e espaciais dos textos. A complexidade dos fatores que contribuem para a marcação deítica torna a tarefa de formalização e anotação automática morosa e complexa. Embora existam já algumas ferramentas[15] que permitam analisar a temporalidade em relação a um determinado momento

[15] Costa & Branco (2012); Hagège, Baptista, & Mamede (2010).

estabelecido, estas ferramentas ainda não conseguem anotar a *deixis*, pelo que a sua identificação e anotação tem de ser feita manualmente. Para este trabalho, a *deixis* interessa-nos em dois momentos: o primeiro, na primeira fase da nossa análise, quando analisarmos as unidades linguísticas que configuram os tipos discursivos, uma vez que a análise das unidades linguísticas será orientada para a identificação dos elementos deíticos presentes nos *corpora*. No segundo momento, recorreremos a estas noções quando as aplicarmos, na segunda fase da análise, na identificação dos tipos discursivos, uma vez que eles se baseiam, na implicação e autonomia do agente, e na convergência ou divergência temporal do enunciado, identificável através da *deixis*.

2. O comentário: caracterização diacrónica

O comentário tem uma longa tradição na atividade académica, enquanto exercício exegético. A sua génese confunde-se com o surgimento da atividade da escrita e da leitura, e a sua prática estendeu-se às mais variadas formas de comunicação na atualidade, criando uma multiplicidade de classes de discurso, em que as características se sobrepõem sem conseguirmos delimitar fronteiras nítidas que estabeleçam a identidade. Importa, por isso, tentar estabelecer uma genealogia do comentário para percebermos (i) se os tipos de comentários que encontramos atualmente têm um tronco comum, de uma perspetiva genológica, e (ii) perceber em que espaços se move esta prática, para determinar o contexto em que é produzido. Para responder a estas questões vamos observar este fenómeno diacronicamente, observando as origens do comentário e de que forma as suas funções se autonomizaram.

2.1. *Hypomnèma*

A palavra grega *hypomnèma*, que significa literalmente "memorando", teve a sua tradução no latim com o termo *commentarium*. Se observarmos o verbete de *commentarius* no dicionário Gaffiot[16] notamos que este se divide em duas aceções, uma com um significado geral: *1 [en gén.] mémorial, recueil de notes, mémoire, aide-mémoire (...);* e um conjunto de significados, agrupados numa aceção mais específica:

> *2 [en part.]*
>
> *a) recueil de notes, journal, registre, archives de magistrats : (...) ;*
>
> *b) Commentarii Cæsaris CIC. Br. 262, les Commentaires de César ;*
>
> *c) brouillon, projet de discours : QUINT. 10, 7, 30 ;*
>
> *d) procès-verbaux d'une assemblée, d'un tribunal : CIC. Verr. 2, 5, 54 ; TAC. Ann. 6, 47 ;*
>
> *e) commentaire, explication d'un auteur : GELL. 2, 6, 1 ;*
>
> *f) cahier de notes [d'un élève] : QUINT. 3, 6, 59.*

[16] https://gaffiot.org/38642. Consultado em 9 de fevereiro de 2023.

Se a significação mais geral do termo, correspondente à aceção 1, associa a palavra comentário à anotação da memória, os significados específicos, embora ainda associados ao registo da memória, têm uma função diferente que se liga à atividade em que são produzidos. Outro aspeto que é importante salientar nesta lista de significados é que há uma distinção em relação ao objeto comentado: encontramos, por exemplo, na alínea *b)* uma referência aos Comentários de César (*De Bellum Gallica)* onde o autor relata as operações militares durante as Guerras da Gália, formando um género literário próprio (Lohfink, 1973: 5), ou na alínea *e)* o comentário enquanto explicação de um autor. A palavra *hypomnèma* veiculou, desde o início, vários significados:

> *le terme [*hypomnèma*] évoque des notes destinées à éclairer des points particuliers, plutôt qu'une œuvre qui exposerait, à nouveau frais, l'ensemble du message véhiculé par le texte de référence*. (Rico, 2003: 5)

Isto significa que, à relação entre comentário e memória, adicionamos também conceitos como esclarecimento, exposição e texto comentado. À medida que as funções do comentário se complexificam e os objetivos do mesmo se expandem, a estrutura também adquire formas específicas. Termos como glosa, escólio, ou marginália emergem como noções de diferentes formas de anotar que, se contêm especificidades (estes termos sobrevivem em atividades especializadas), também se sobrepõem, com frequência, por terem pontos em comum. No entanto, existem três termos que consideramos importante distinguir: o glossário, a glosa e o escólio. Os glossários são principalmente explicações de palavras que foram desenvolvidos na Antiguidade para explicar palavras obsoletas (Lohfink, 1973: 6) ao passo que a glosa são pequenas notas na língua do texto, marginais ou interlineares, tendo como objetivo explicar palavras ou passagens obscuras ou de difícil entendimento (Calabrese, 2019: 9). Relativamente ao escólio, as opiniões divergem: enquanto alguma literatura especializada estabelece a distinção entre glosa e escólio pelo lugar que ocupa na *codex* (cf. Dickey, 2007), outros autores estabelecem uma distinção pela complexidade do comentário:

> *Ein antikes Scholion erklärt zwar den Text vielschichtiger als eine Glosse. Scholien kommentieren aber nur einzelne Sätze oder Textabschnitte, hingegen nicht fortlaufend ein ganzes Werk*[17]. (Lohfink, 1973: 6)

A diferença entre glossário, glosa e escólio é ténue uma vez que as funções se sobrepõem e, frequentemente, é difícil estabelecer uma linha que separe estas práticas textuais. Mas é importante notar que todas estas noções ainda contêm não só a ideia de memória, mas também de pensamento, do pensamento sob a forma de anotação. Podemos mesmo acrescentar que o comentário foi, antes de ser uma prática discursiva, uma ferramenta do pensamento: na Antiguidade, o comentário tem origem numa prática de trabalho, mais do que num género. Como sintetiza (Calabrese, 2019: 9):

> *Autrement dit, la catégorie des hypomnemata comprend un large spectre de significations, de modèles et de types de textes (Dorandi 2000 : 27) qui correspondent à la pratique de commenter un texte préexistant.*

Embora, como aponta Calabrese (2019), a prática do comentário se expresse de diferentes modos e em função de diversos "tipos" de texto, o facto de comentar algo que pré-existe a si, e ao qual está umbilicalmente ligado, é o elemento comum às diversas práticas. A pergunta que podemos colocar neste momento é se a relação, ou autonomia, entre o comentário e o texto comentado é idêntico em todas as práticas. A ideia de autonomia textual é complexa e filia-se em várias tradições escolares ocidentais, com origem em escolas da antiguidade, tradições rabínicas e cristãs. Estas últimas são particularmente importantes porque, como explica Rico (2003), o tipo de comentário varia de acordo com a função / objetivo dessas tradições. Se na tradição rabínica, o comentário tem duas funções - o esclarecimento (função semântica) e a atualização do texto religioso (função pragmática) - em que está subjacente uma devoção do texto escrito (enquanto palavra de Deus, e, portanto, perfeita), na tradição cristã, a Bíblia tem um único significado que o comentário tenta fixar, impondo uma leitura universal dentro de uma tradição homilética:

[17] [Tradução de Estela Tschernutter: "*Um antigo escólio esclarece o texto de forma mais complexa. Mas, em contrapartida, os escólios comentam apenas uma única frase ou uma passagem e não uma obra completa de forma contínua*"].

> *le commentaire chrétien vaut dès lors pour lui-même, comme un discours autonome et non pas comme une annotation du texte biblique. De la Bible, il dégage des péricopes qu'il réexpose en les actualisant*. (Rico, 2003: 20)

Esta afirmação mostra-nos que o comentário de tradição cristã parte de uma noção de autonomia que se inscreve numa leitura do significado geral do texto (lembremos, por exemplo, que a tradução integral da Bíblia para línguas vernáculas foi sempre desencorajada, circulando apenas partes traduzidas da *Vulgata Latina*), enquanto o comentário de tradição rabínica parte de uma leitura sincrónica do significado da palavra. Partindo destas noções, poderíamos ainda questionarmo-nos sobre a importância do comentário oral e do comentário escrito e se existem áreas em que o comentário de tradição rabínica e o comentário de tradição cristã se sobrepõem. Não estando no âmbito deste trabalho fazer um estudo detalhado sobre esta matéria, referimos estas questões para mostrar que existem múltiplas abordagens ao estudo genológico do comentário e que existem variantes que concorrem e coocorrem para a indefinição do comentário enquanto género.

A autonomia do comentário em relação ao texto comentado é, como acabámos de ver, uma das características que permite fazer a filiação do comentário, sobretudo em tradições mais pretéritas. Vamos assumir que, para este trabalho, a autonomia é uma propriedade gradiente que parte do polo dependente (grau zero de autonomia) em direção ao polo independente (grau mais elevado de autonomia). No grau zero de autonomia colocaríamos, por exemplo, os glossários (marginal ou entrelinhas), seguido das glosas, escólios e sucessivamente até atingir o polo oposto que seria o comentário enquanto texto, com uma autoridade autoral e editorial que permite a sua leitura sem a presença[18] do texto comentado.

> *Le commentaire peut prendre la forme d'un texte à part entière (décliné en plusieurs genres : littéraire, philosophique, historique, biblique…) ou bien d'une pratique générale d'annotation des textes, raison pour laquelle le terme est souvent employé comme un hypéronyme de toutes les formes de l'activité de gloser*. (Calabrese, 2019: 10)

[18] A questão da presença ou ausência do texto comentado liga-se fortemente à noção de intertextualidade que não vamos abordar aqui.

A noção de comentário funciona como hiperónimo de um conjunto variado de práticas textuais, com graus diversos de autonomia em relação ao objeto comentado.

2.2. A página enquanto espaço enunciativo

Como já referimos no início desta secção, a noção de comentário tem origem na ideia de memorando, à qual se foram juntando outras noções como esclarecimento e exposição, e que complexificaram a disposição textual, adquirindo funções e estruturas próprias. Vimos também que, antes de ser um género, o comentário foi uma prática de trabalho, que a passagem do ato de anotar esteve, entre outros aspetos, ligada ao desenvolvimento dos suportes físicos, e que estes modificaram a forma como o leitor se relaciona com o texto. Mas de que forma a página, enquanto lugar de criação intelectual, passou a um espaço enunciativo?

O surgimento do *codex* enquanto suporte de escrita, além das vantagens práticas (espaço ocupado, manuseamento, hierarquização da informação), permitiu organizar de forma mais racional o texto e construir a página enquanto espaço enunciativo. Esta passagem faz nascer intelectualmente a página, que surge como um elo entre a prática da escrita e o texto. Assim, o enquadramento proporcionado pela página torna-se uma ferramenta (intelectual) e o dispositivo gráfico transforma-se numa unidade linguística e permite, a partir deste momento, a semantização do espaço gráfico. O enquadramento gráfico do texto funciona como um elemento operatório que orienta e organiza a informação e a sua leitura, permitindo a construção de uma interpretação crítica do texto:

> *The specific proposition is that writing, and more especially alphabetic literacy, made it possible to scrutinise discourse in a different kind of way by giving oral communication a semi-permanent form; this scrutiny favoured the increase in scope of critical activity, and hence of rationality, scepticism, and logic to resurrect memories of those questionable dichotomies.* (Goody, 1977: 37)

De acordo com Goody, a projeção gráfica permite a agenciação de significados num espaço bidimensional, e a disposição dos signos linguísticos numa determinada superfície permite organizá-los e hierarquizá-los, numa relação de sentido, favorecendo o espírito de reflexão e o surgimento de um metadiscurso.

> *L'apparition du codex vers la fin du Ier s. de notre ère va cependant rendre possible une véritable révolution dans la présentation matérielle de ces commentaires. De nombreux spécialistes ont souligné le lien entre l'ampleur des marges du codex de l'Antiquité tardive et le développement des scholies paratextuelles.* (Rico, 2003: 11)

A evolução dos suportes físicos favoreceu a evolução das práticas do comentário, e o comentário tornou-se uma ferramenta de gestão do texto comentado: se por um lado alguns tipos de comentários (como os marginais e interlineares) permitem esclarecer o texto em que surgem, por outro, a seleção de excertos ou o uso do texto integral permitiu, como referimos na secção anterior, estabelecer leituras canónicas e adquirir um grau de autonomia que os primeiros não conseguiram.

2.3. O comentário enquanto género.

Na secção anterior vimos que a função do comentário não se esgotou no memorando ou como um orientador da leitura. Vimos também que o seu desenvolvimento está ligado a vários fatores que vão da ferramenta hermenêutica à própria história da leitura e da relação do leitor com o objeto livro, e que inicialmente não constituía um género com marcas definidas, mas uma prática de trabalho. A questão que agora nos surge é se houve algum momento em que o comentário adquiriu o estatuto de género textual.

Lohfink (1973) estabelece três critérios para a definição do comentário enquanto género, dentro da tradição exegética. O primeiro critério é a interpretação contínua do texto, que excluiu do género comentário as glosas e os escólios. O segundo refere-se à adequação da interpretação relativamente ao texto interpretado, isto é, o comentário não traz nenhum material estranho para o texto comentado, apenas a interpretação, excluindo os comentários de tradição rabínica. O terceiro critério prende-se com a intenção da linguagem e estipula que o comentário tem uma intenção argumentativa e não apelativa, o que exclui os textos de tradição homilética. São critérios discutíveis e o autor assume que não dão conta da multiplicidade de comentários com características comuns ou "aparentados" entre si:

> *Denn das ist ein zweites Gesetz einer allgemeinen Gattungs- geschichte: Gattungen existieren nie als chemisch reines Substrat, sondern stets als konkrete*

> *sprachliche Formen, die fast immer Affinität zu verwandten Gattungen aufweisen oder die sogar dicht an der Grenze zu anderen Gattun- gen stehen.* [19] (Lohfink, 1973: 9)

Consideramos, por isso, que no passado o comentário só pode ser considerado um género textual quando se estabelece enquanto prática textual com uma função explicativa ou pedagógica, e que esta passagem ocorre por volta do séc. V (Rico, 2003: 7) . De acordo com o autor, é neste momento que a estrutura dos textos explicativos se começa a fixar nas escolas:

> *Ce philosophe [Proclus] établit en effet un modèle d'introduction aux commentaires des textes d'Aristote et de Platon, dont l'étude constituait la base du cycle des études néoplatoniciennes. Pourtant, dès le IIe siècle de notre ère, on voit apparaître les principales étapes de ces introductions, malgré certaines fluctuations dans l'ordre et la terminologie exacte des éléments:* skopos *(but de l'œuvre),* chrèsimon *(utilité),* gnèsion *(authenticité),* taxis *(place de l'ouvrage dans l'ordre de lecture),* aitia tès epigraphès *(raison du titre),* diairesis *(division des livres). Une fois reconnu le dessein général de l'œuvre, l'auteur d'un commentaire devait en effet signaler la place de l'ouvrage étudié dans l'ensemble du corpus d'un philosophe et dégager les divisions du texte, avant de procéder au commentaire proprement dit.* (Rico, 2003: 7, 8)

Significa, portanto, que (i) a prática do comentário passou de uma função mnemónica para uma função explicativa, que (ii) já está associada ao estabelecimento da ideia de um cânone, e que (iii) o comentário adquiriu um método de redação com um plano de texto [20] definido. Estes três aspetos contribuíram para que o comentário adquirisse um caráter normativo e disciplinador da interpretação dos textos, desenvolvendo-se como um género dentro das disciplinas em que é praticado.

> *Autrement dit, le commentaire instaure une manière de lire restrictive tout en ayant l'air de répéter le texte premier. Cette fonction du commentaire comme*

[19] [Pois essa é a segunda lei de uma História Geral dos Géneros: os géneros não existem nunca como um substrato quimicamente puro, mas sempre como formas linguísticas concretas, que apresentam quase sempre afinidade com outros géneros análogos ou que estão até bem próximos da fronteira com outros géneros.]

[20] A noção de plano de texto será abordada, posteriormente, no quadro do ISD.

> *genre fait écho à la fonction de contrôle social de l'écriture comme pratique réservée à une élite (...)* (Calabrese, 2019: 15)

Esta função de controlo social da escrita cria, também, a figura do autor enquanto autoridade material e moral do comentário, e que irá perdurar até meados do séc. XV quando começam as edições dos autores clássicos (destinados ao ensino) nas línguas originais e expurgadas dos comentários. O fim do comentário escolástico e dos seus autores enquanto figuras de autoridade singulares dará origem a autoridades múltiplas:

> *Os professores da escola latina de Sélestat transmitiam preceitos ortodoxos que implicavam a existência de uma leitura comum «correcta», mas também ofereciam aos estudantes uma perspectiva humanista mais vasta e mais pessoal; os estudantes acabaram por reagir inscrevendo a leitura no âmbito do seu universo íntimo e da sua experiência pessoal e afirmando a sua autoridade de leitores individuais sobre todos os textos.* (Manguel, 2020: 123)

De uma ortodoxia da leitura, reservada a uma elite e praticada em público, passámos à multiplicidade de interpretações individuais em que cada leitor é um "autor potencial" (Calabrese, 2019: 16), contribuindo para a autonomização editorial e autoral do comentário, numa prática que se estendeu à época contemporânea.

Em resumo

- O comentário teve, na sua origem, uma função mnemónica enquanto ferramenta de trabalho.
- A evolução do comentário teve dois eixos fundamentais: um eixo hermenêutico, associado à função, e um eixo antropológico, ligado ao desenvolvimento dos suportes materiais utilizados pelo Homem enquanto espaços enunciativos.
- À medida que as funções e o formato (disposição textual) se tornaram mais específicos, as diversas práticas do comentário adquiriram graus de autonomia diferentes.

Partindo deste último ponto, consideramos que o comentário era, na sua origem, uma etiqueta para várias práticas textuais com funções diversas que frequentemente se

sobrepunham, evoluindo na tradição escolástica para um género restrito praticado em diferentes disciplinas (religioso, filosófico, filológico e literário).

Na secção seguinte, faremos uma síntese da passagem do comentário, restrito a áreas do conhecimento específicas, à sua democratização e ocupação de um lugar de destaque na comunicação social.

3. O comentário contemporâneo: caracterização sincrónica

Como vimos na secção anterior, o comentário teve origem numa prática de trabalho, com uma função mnemónica e com uma estrutura pouco definida que evoluiu conjuntamente com o desenvolvimento técnico, académico e filosófico, para um género com uma função normativa. Mas na sociedade contemporânea, o comentário adquiriu um estatuto omnipresente nos meios de comunicação, principalmente nas redes sociais e na atividade jornalística. Se, no passado, o comentário era uma prática que se definia pela atividade em que era exercido e pela função normativa que veiculava, no presente, também o objeto do comentário se modificou.

> *De acordo com o exposto, fica claro que a noção de comentário evoluiu ao longo do tempo, quer pelo contexto de uso (inicialmente mais académico e atualmente mais quotidiano e digital), quer pelo foco (outrora sobretudo em textos e recentemente em objetos semióticos de ordem diversa como textos, vídeos, fotografias).* (Gonçalves & Carrilho, 2020: 193)

A mudança do foco do texto para outros objetos semióticos fez desaparecer um dos elementos estabilizadores do comentário enquanto género, e fez surgir um novo conjunto de práticas discursivas, algumas mais próximas do diálogo do que do comentário (Calabrese: 2019; Gonçalves & Carrilho: 2020). Estas mudanças tiveram impacto na forma como a prática do comentário ocorre atualmente:

- A aproximação ao género conversacional alterou a relação entre o produtor do texto e o leitor, passando de uma relação global e vertical (de autoridade) para uma relação local e horizontal, onde ocorre uma negociação entre os intervenientes (Maingueneau, 2007: 31).
- As redes sociais são construídas para que haja uma reação/interação entre o produtor e o leitor (Calabrese, 2019: 18), ocorrendo uma redistribuição dos papéis discursivos. Assim, o comentário dá lugar ao diálogo (confronto), e a função explicativa dá lugar à opinião, passando do discurso erudito[21] para o discurso do senso comum.

[21] Entendemos o discurso erudito como um discurso construído sobre figuras de autoridade, e que é validado academicamente.

Não estando no âmbito deste trabalho fazer um estudo aprofundado sobre as diversas práticas do comentário nos meios digitais (cf. Valentim & Gonçalves (2021) e Calabrese & Jenard (2018)), a sua referência é necessária para mostrar como, mais uma vez, a democratização da leitura e da escrita conduziu à alteração da noção de um género comentário. Esta alteração ocorreu também na imprensa escrita onde o comentário é utilizado como etiqueta para diversas práticas textuais.

> *Dificilmente poderemos assumir que se trata, em todos estes casos, de um mesmo género de texto, mesmo que a etiqueta usada para os referir seja a mesma. Apesar de partilharem, certamente, algumas características, comentários associados a diferentes práticas sociais (coletivas) corresponderão a diferentes géneros de texto (...).* (Coutinho, 2019: 111)

Partindo desta citação, há duas questões que se impõem: (i) o que é que sobrevive do género autoral do comentário, (ii) como é que ele se manifesta na contemporaneidade na imprensa escrita (impressa ou digital)?

3.1. O comentário na imprensa escrita.

O comentário (entendido como um género autoral, com uma função explicativa) convive nos meios de comunicação escritos com outros textos de géneros vizinhos, como a opinião e o editorial, e a flutuação entre etiquetas e designações é frequente, como mostram alguns estudos (Adam, 1997; Laidouni, 2019; Gonçalves & Carrilho, 2020). No entanto, não podemos esquecer que as etiquetas[22] são atribuídas pelos profissionais da atividade em que se inscrevem, e que estas emergem *dans le cadre des interactions verbales propres à une formation discursive donnée (un journal ou un type de presse, et plus largement un média donné : radio, télévision, presse écrite, édition ou cinéma)* (Adam, 1997: 7). Isto significa que a genealogia dos textos de imprensa está codificada e é transmitida profissionalmente através de manuais de escrita jornalística, como mostram Adam (1997) e Laidouni (2019), e que as suas definições e critérios

[22] As etiquetas, que iremos abordar brevemente no enquadramento teórico do ISD, fazem parte daquilo que Adam (1997: 5) denomina de peritexto, e que é o conjunto de elementos que gravitam em torno do texto, como os títulos, subtítulos ou rubricas. A sua importância para a questão dos géneros redaccionais, explica o autor, reside não tanto no seu atributo genérico mas na forma como os géneros selecionam determinados elementos peritextuais e excluem outros.

convergem em maior ou menor grau, afastando a ideia de anarquia de géneros. Laidouni (2019: 40) dá como exemplo desta codificação o caráter repetitivo da transmissão da informação em cada número ou edição. Adam (1997) parte desta abordagem para analisar as unidades linguísticas, textuais e composicionais, para descrever as categorias da imprensa escrita[23] os géneros discursivos na imprensa escrita.

3.2. Os géneros do jornalismo através dos manuais

Adam (1997) analisou um conjunto de manuais de jornalismo e situou os géneros jornalísticos dentro de cada uma das interações, específicas das formações discursivas. No trabalho de Broucker (1995) sobressaem dois grandes géneros na imprensa: o género informação e o género comentário. De acordo com Adam, Broucker utiliza três critérios para separar estes géneros:

- critério semântico: o sujeito;
- critério argumentativo e pragmático: a intenção do texto que pode ser informativa ou explicativa:
- critério enunciativo: a posição do jornalista relativamente ao seu discurso.

Estes critérios permitiriam colocar todos os géneros de texto da imprensa dentro do género informação ou do género comentário. Uma observação que podemos fazer neste momento é se aquilo que Broucker designa por géneros informação / comentário serão verdadeiramente géneros textuais. Adam (1997: 9), considera que se trata mais de posições enunciativas do que propriamente de géneros, e organiza os textos que se inscrevem na atividade jornalística em torno destas duas posições enunciativas: o polo informação corresponde ao nível mais autónomo em termos de posição enunciativa, onde as práticas textuais que se enquadram neste polo são, por exemplo, "Dépêche", "Brève" e "Filet": O termo "dépêche" refere-se a uma comunicação das agências de notícia, concisa mas com muitos dados relevantes, sendo equivalente ao "despacho" ou "notícia de agência" em português. O "brève" é um texto curto, muitas vezes com apenas algumas linhas, que resume um evento ou facto de maneira direta, sendo

[23] "*catégories de la presse écrite*" ou "*unités redactionelles*" (Adam, 1997).

conhecida como "nota" ou "nota curta" em português. Já o "filet" é um texto breve, mas um pouco mais detalhado que a "brève", com foco na concisão e clareza, e pode ser comparado ao "resumo informativo". São tipos de textos comuns em agências de notícias e meios de comunicação que precisam de divulgar informações de forma rápida.

	De Broucker	Martin-Lagardette	Antoine, Dumont, Grevisse, Marion, Ringlet	Montant
Pôle distance-information				
1	Dépêche			
2	Brève	Brève		Brève
3	Filet	Filet		Filet
4	Communiqué		Communiqué	
5	Texte d'auteur			
6	Revue de presse			Revue de presse
7	Information-service			
8		Résumé de rapport		
9	Compte-rendu	Compte-rendu	Conférence de presse	Compte-rendu
10	Enquête	Enquête	Enquête	Enquête
11	Reportage	Reportage	Reportage	Reportage
12			Fait divers	
13			Papier d'ambiance-observation	
14	Interview	Interview	Interview	Interview
14a	Interview-enquête	Interview-information		
14b	Interview-reportage	Interview-information		
14c	Interview-rencontre			
14d	Interview-documentaire			
14e	Interview-sondage	Interview-express		
14f		Interview-d'opinion		
14g		Interview-portrait		
14h			Interview-interrogatoire	
14i			Interview-conversation	
14j			Interview-récit	
15	Portrait	Portrait (profil)		Portrait
16		Article de commentaire		Article d'analyse
16a	Commentaire explicatif			
16b	Commentaire-traduction			
16c	Commentaire interpréta-tif			
16d	Commentaire expressif			
17	Éditorial	Éditorial	Éditorial	Éditorial
18	Tribune	Tribune libre		« Billet »
19	Courrier des lecteurs	Courrier des lecteurs		
20	Papier d'expert			
20a		Critique	Critique	Critique
21	Billet	Billet	Billet d'humeur	« Humeur »
22	Caricature			
23	Chronique	Chronique	Chronique judiciaire	Chronique
24		Écho	Écho	Écho et ragot
Pôle implication-commentaire				

Figura 5: Organização dos géneros jornalísticos em polos; extraído de Adam (1997: 11)

No polo oposto, o polo implicação-comentário encontramos práticas textuais como a caricatura e a crónica, sendo esta última um texto que mistura análise, opinião ou observações pessoais. Note-se, ainda, que o comentário, embora mais próximo deste polo implicação, surge quase a meio da tabela, dividido em quatro tipologias: "Commentaire explicatif", "Commentaire-traduction", "Commentaire interprétatif" e "Commentaire expressif", sugerindo diversos graus de implicação para uma prática textual que funciona como uma espécie de hiperónimo.

Este quadro traz também para a discussão as noções de distanciamento (ou autonomia) e implicação, que iremos discutir mais profundamente quando as definirmos dentro do quadro do ISD. No entanto, esta organização em polos gradientes continua a não dar conta da multiplicidade de etiquetas que podemos encontrar na imprensa escrita:

> *Cette complexité et les différences s'expliquent par des croisements de critères qui vont des choix stylistiques micro-linguistiques aux intentions communicatives, en passant par la position énonciative du locuteur et le contenu des articles. Selon que tel ou tel critère est mis en avant, les catégories bougent sensiblement.* (Adam, 1997: 11)

Partindo desta complexidade, o autor irá teorizar sobre a noção de género e de planos de organização textual, e que definimos na primeira parte deste trabalho, dentro do enquadramento teórico.

Em resumo:

- O comentário contemporâneo assumiu diversas formas e o seu objeto multiplicou-se noutros objetos semióticos.
- Nas plataformas digitais, com destaque para as redes sociais, aproximou-se do género dialogal, em que é esperada uma reação e interação entre o comentário e os leitores.
- A aproximação ao género dialogal conduziu à redistribuição dos papéis discursivos.

- O comentário, enquanto género autoral, sobrevive em várias áreas da imprensa escrita, mas os critérios pelos quais são definidos pelos manuais da especialidade, sejam eles composicionais (como as etiquetas) ou enunciativos, variam sem encontrarmos uma lógica que os una.

Serviu este capítulo para fazer uma caracterização sincrónica do comentário, mostrando como ele sobrevive atualmente, sobretudo nos meios de comunicação social (e outras atividades específicas) mas serviu, também, para mostrar que o comentário é uma prática textual cuja tipificação ou caracterização depende de vários fatores. No próximo capítulo faremos uma descrição da metodologia para efetuar a nossa análise ao comentário.

4. Metodologia

4.1. Introdução

Nos capítulos anteriores, estabelecemos o quadro teórico e as noções relevantes para a análise do *corpus* textual. No presente capítulo apresentaremos a metodologia e os *subcorpora* utilizados no nosso trabalho.

A linguística de *corpus* tem assumido um papel relevante nos estudos da linguística do texto. McEnery & Wilson (2001) mostra alguns exemplos de como o uso de *corpora* em estudos de texto, especialmente na área da estilística e da linguística do texto, pode ser útil para a identificação das características do género textual. O autor refere também que os *corpora* podem igualmente ser usados para desafiar abordagens tradicionais de tipologia textual, como no trabalho de Biber (1988), que utilizou a análise fatorial para identificar dimensões de variação linguística em géneros diferentes. Tal abordagem permite uma comparação mais ampla e empiricamente fundamentada da variação linguística entre os géneros, em detrimento de análises localizadas a características isoladas. McEnery reforça ainda a importância de se considerar as semelhanças e diferenças gerais entre os géneros, em vez de focar as variações de características individuais, e a necessidade de construir *corpora* mais representativos, baseados numa perspetiva interna da língua.

A representatividade é outro fator importante quando se trabalha com *corpora*, e que também é abordada por McEnery (2001: 77-81). De acordo com o autor, há alguns aspetos relevantes que devem ser tidos em conta para garantir a representatividade dos dados, e que passamos a elencar:

(i) Definir claramente os limites da população de interesse: Antes de realizar a amostragem, é fundamental definir com precisão o que constitui a população a ser estudada. Definir, por exemplo, os textos publicados num determinado período.

(ii) Escolher uma abordagem de amostragem adequada: Para textos publicados, pode-se utilizar um índice bibliográfico abrangente como base para definir a amostra.

(iii) Considerar métodos de amostragem demográfica para linguagem informal: Para recolher dados de linguagem informal, como conversas ou correspondências

privadas, que não estão formalmente indexados, recomenda-se o uso de uma amostragem demográfica, baseada em características como idade, sexo, classe social e região.

(iv) Complementar a amostragem demográfica com uma abordagem contextual: Reconhecendo que a amostragem demográfica pode não cobrir todos os tipos importantes de linguagem, é necessário complementar com uma amostragem dos tipos de atividades linguísticas contextualmente relevantes, que não seriam captados pela amostragem demográfica.

(v) Definir a estrutura hierárquica da amostra (estratificação): Antes de realizar a amostragem, deve-se proceder ao mapeamento da estrutura hierárquica da amostra, como por exemplo diferentes géneros (por exemplo, ficção científica, reportagens de jornais, escrita jurídica, etc.), para ajudar a garantir que a categoria relevante esteja adequadamente representada.

(vi) Utilizar amostragem estratificada em vez de amostragem puramente probabilística: o conceito de estratificação refere-se à divisão dos textos em diferentes subgrupos ou categorias antes de se proceder à amostragem. A estratificação implica identificar e separar os textos de acordo com características distintas, como diferentes géneros ou canais de comunicação, para garantir que cada uma dessas categorias seja devidamente representada na amostra. A amostragem estratificada é muitas vezes mais representativa do que a amostragem probabilística pura, pois permite fazer uma representação adequada de cada estrato ou género.

(vii) Determinar um tamanho adequado para as amostras: O tamanho da amostra deve ser determinado de acordo com a distribuição das características linguísticas dentro da população. Itens mais frequentes exigem amostras menores, enquanto itens mais raros requerem amostras maiores para garantir a sua representação.

(viii) Usar as características com maior variabilidade para calcular o tamanho das amostras: Para garantir que as amostras sejam representativas, recomenda-se calcular o tamanho das amostras com base nas características linguísticas que mostram mais

variação na população. Deste modo, assegura-se que os textos com características mais raras sejam também adequadamente representados no *corpus*.

(ix) Aplicar medidas de dispersão para melhorar a representatividade: Medidas de dispersão podem ser utilizadas para garantir que a ocorrência de um item ou fenómeno linguístico esteja bem distribuída no *corpus*, em vez de estar concentrado numa pequena parte. Só assim é possível medir se uma palavra ou expressão é frequente dentro de um género ou do *corpus* como um todo.

(x) Adotar uma abordagem estatística rigorosa: A aplicação consistente de métodos estatísticos ajuda a garantir que o *corpus* seja o mais representativo possível, levando em conta as limitações práticas do processo de construção do *corpus*.

Os aspetos elencados pelo autor não se aplicam na totalidade ao nosso trabalho, uma vez que o foco desta lista é assegurar uma amostra adequada para fenómenos linguísticos do tipo micro, e a nossa abordagem, como já referimos, pretende analisar o texto como uma unidade. No entanto, oferece pistas importantes para a forma como devemos constituir os nossos *corpora* de análise: definir claramente a amostra, complementar com uma abordagem contextual e mapear a amostra para garantir que as categorias em análise estejam adequadamente representadas.

O uso de *corpora* para a análise de texto não é novo, e os trabalhos de análise de Bronckart foram desenvolvidos a partir de um *corpus* de textos empíricos "en situation naturelle ou expérimentale" (Bronckart, 1997: 80), que foi objeto de uma análise[24] qualitativa e quantitativa. Também François Rastier tem desenvolvido a sua investigação em torno da semântica textual, evidenciado a importância dos estudos linguísticos baseados em *corpora*:

> *La linguistique de corpus pourvoit ainsi la linguistique d'un domaine où elle peut élaborer des instruments et définir une méthode expérimentale propre: elle ouvre aussi des champs d'application nouveaux et engage un mode spécifique d'articulation entre théorie et pratique.* (Rastier, 2011)

[24] Descrita em pormenor em Bronckart (1988).

A nossa análise desenvolve-se em três fases, cada uma com o seu *corpus* de análise. A opção pela divisão em vários momentos justifica-se pelo tipo de fenómeno que vamos analisar, em que cada uma dessas divisões, bem como a estratificação do próprio *corpus* – isto é, a divisão que vamos fazer em *subcorpora* para separar os textos em categorias distintas – se liga à relevância quantitativa e qualitativa dos dados em cada uma destas etapas de análise. A primeira fase contempla as unidades linguísticas e tem como objetivo observar as variações microlinguísticas através da análise da anotação morfossintática. Na segunda fase da análise serão analisados os tipos discursivos presentes nos textos. Na terceira fase, procuraremos a identidade do género comentário não só através dos TD (que configuram unidades linguísticas específicas), mas também de elementos praxeológicos e contextuais, como a atividade e o tema. Definimos, em termos metodológicos, que para cada umas das três fases do trabalho deveriam ser utilizados *corpora* diferentes, uma vez que os objetivos e a validação dos dados são diferentes:

- Fase 1: CETEM, COMENTA2 e COMJUR
- Fase 2: COMENTA2 e CETEM2
- Fase 3: COMENTA2 e CETEM2

O COMJUR é analisado, na primeira fase, de forma independente como um *subcorpus*, e é integrado no COMENTA2 nas fases 2 e 3.

Na secção seguinte, faremos uma descrição detalhada dos *corpora* utilizados em cada um dos momentos de análise.

4.2. Descrição dos *corpora*

4.2.1. Origem dos Dados

Para a primeira fase, utilizámos o *corpus* CETEMPúblico[25] que inclui o texto de cerca de 2600 edições jornal diário *Público*, entre os anos de 1991 e 1998. De acordo com Rocha & Santos (2000), o *corpus* é composto por 1.567.625 extratos de texto em português europeu, com classes de texto variados. Inicialmente, o jornal possuía uma classificação própria que se alinhava com as diferentes seções do periódico (cultura,

[25] Disponível no endereço: https://www.linguateca.pt/CETEMPublico/

política, economia, entre outros). No entanto, no tratamento do *corpus* para ser disponibilizado, a equipa que fez o tratamento dos textos optou por uma reclassificação do conteúdo, fundamentada na premissa de que, para fins de processamento de linguagem natural (PLN), a classificação mais significativa reside no conteúdo temático ou no estilo textual, em detrimento da formatação ou da posição física do texto no jornal[26].

Utilizámos também, para esta primeira fase, o *corpus*[27] G&T.Comenta recolhido no âmbito das atividades do grupo Gramática e Texto, integrados no CoRus - Projeto Estratégico 2015-2020, desenvolvido pelo Centro de Linguística da Universidade Nova de Lisboa (CLUNL). Este *corpus* tem características mais heterogéneas porque resulta da recolha de textos com circulação em diversos suportes e de diferentes origens. Ao contrário do CETEMPúblico, este *corpus* não tinha como objetivo inicial o uso para fins de PLN, e pretendia estudar e catalogar o comentário enquanto atividade de linguagem e prática textual. Assim, o parâmetro que presidiu à recolha dos textos foram as "etiquetas" que acompanham os textos recolhidos (Gonçalves & Carrilho, 2020), assumindo que a sua atribuição contenha um grau de arbitrariedade ou subjetividade que pode fazer incluir textos que não são verdadeiramente comentários e excluindo outros que o são. Deste modo, foram recolhidos 791 textos, escritos em português europeu, que contêm as etiquetas "comentário" ou "comentador". Num segundo momento do *corpus* G&T.Comenta, os textos foram catalogados com um conjunto de descritores para que pudessem ser pesquisados e acedidos. Dentro dos diversos descritores utilizados (Gonçalves & Carrilho, 2020), é importante destacarmos a distinção que foi feita entre os tipos de comentário:

(i) Os comentários de leitor: textos feitos em caixas de diálogo.

(ii) Os comentários de utente: textos em sítios que apresentam um produto que é o objeto de um comentário.

[26] A questão da organização da informação/notícias nas diversas seções dos jornais foi abordada em Gonçalves & Magalhães (2019).

[27] Disponível no endereço: https://projetos.dhlab.fcsh.unl.pt/s/GTComenta/page/projeto

(iii) Os comentários de autor: textos produzidos por uma determinada personalidade, em formato de artigo.

Como referimos na terceira parte deste trabalho, a noção de comentário funciona como hiperónimo de um conjunto variado de práticas textuais, com graus diversos de autonomia em relação ao objeto comentado. Na Figura 6 esquematizamos o nosso entendimento relativamente a um comentário que se enquadra no texto autoral e editorial (com autor identificado), em que o grau mais elevado de autonomia se caracteriza pela ausência do objeto comentado no mesmo espaço enunciativo (podendo ser convocado no próprio texto) e a identificação do autor do comentário, enquanto autoridade. No polo oposto, consideramos que o comentário tem menos autonomia quando é exigida a presença desse outro objeto no espaço enunciativo, e a identificação do autor do comentário é opcional.

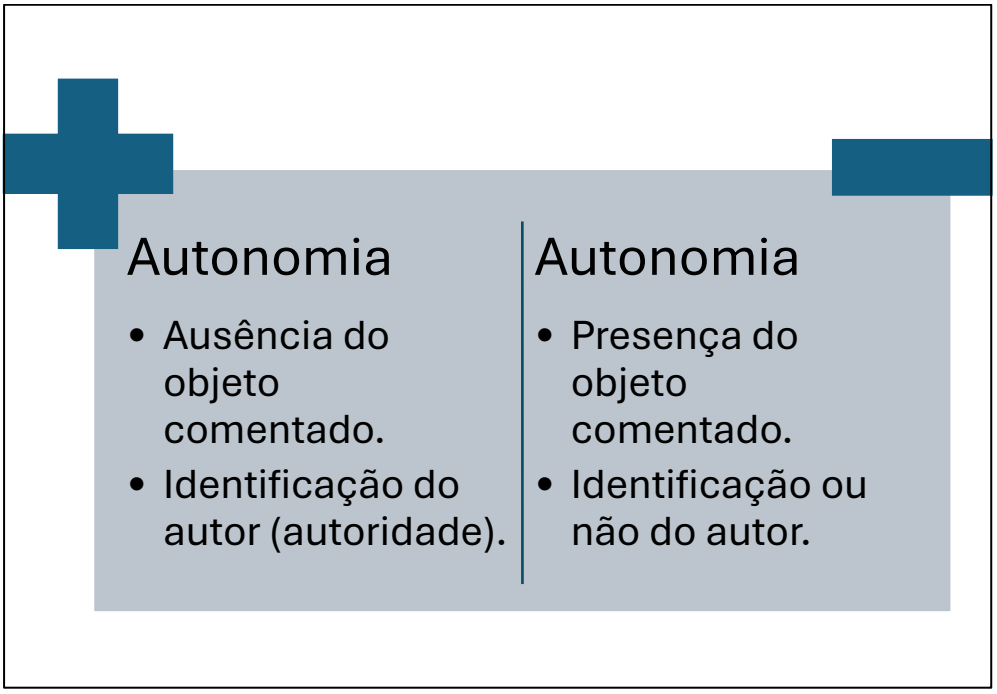


Figura 6: Autonomia do comentário em relação ao objeto comentado.

Foi partindo desta noção que fizemos uma seleção de textos do *corpus* G&T.Comenta para que cumprissem o requisito de serem comentários de autor. Referir-nos-emos a esta seleção de textos por COMENTA2.

Como vimos anteriormente, os textos que fazem parte do *corpus* G&T.Comenta têm origem em diversos suportes e origens, mas ao observarmos o gráfico com a

distribuição dos textos por Atividade, damo-nos conta que ao retirarmos os textos da atividade hoteleira (comentários de utente) e das redes sociais (comentário de leitor), ficam apenas com os comentários da atividade jornalística, académica e literária, sendo estas últimas semelhantes em alguns aspetos (Coutinho, 2019: 111).

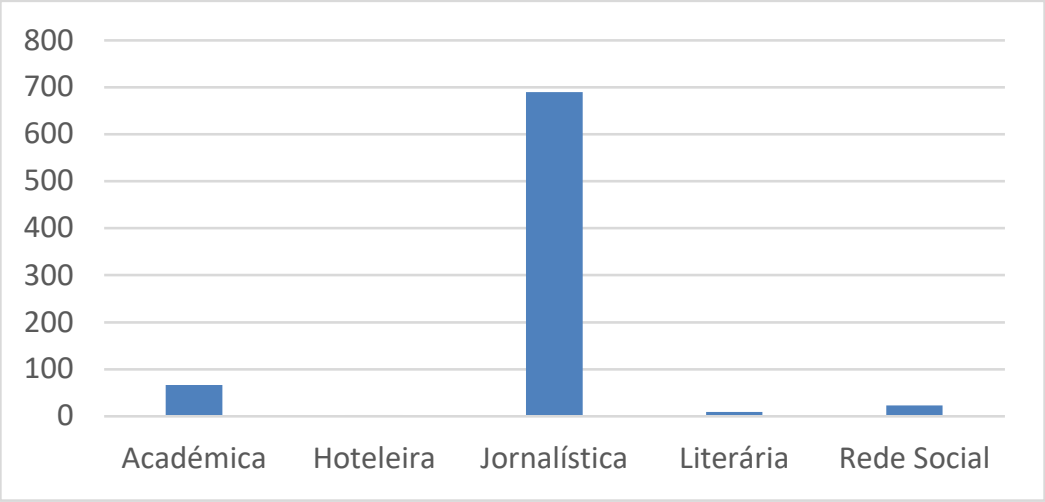


Figura 7: distribuição do Corpus por Atividade, extraído de Gonçalves & Carrilho (2020: 201)

Para aumentar a diversidade de Atividades, adicionámos dois textos produzidos na atividade jurídica, e recolhidos por nós em sites de especialidade. O critério de recolha destes comentários jurídicos foi a presença da etiqueta "comentário", uma vez que não observámos a existência da etiqueta "comentador" nesta área de atividade.

4.2.2 Composição dos *corpora*:

O *corpus* CETEMPublico, versão 1.7, é constituído por cerca de 190 milhões de palavras. Partindo dos números do COMENTA2 e jurídico (doravante designado COMJUR), porque são *corpora* fechados, selecionamos uma parte do CETEM que correspondesse a um número similar de palavras. A opção pelo número de palavras deve-se ao facto de, como explicámos anteriormente, o CETEM ser constituído por segmentos de texto (dois parágrafos, de tamanho variável, do texto original) e não por textos completos. Deste modo, os dados quantitativos aproximados sintetizam-se no quadro seguinte:

	Tokens	Palavras
CETEM	233.016	~186.878
COMENTA	225.135	~180.558
COMJUR	6.884	~5.520
Total	465.035	372.956

Tabela 3: Síntese do número de palavras e tokens do corpus.

Embora os textos jurídicos adicionados surjam aqui separados, uma vez que foi adicionado por nós ao *corpus* COMENTA, consideramos que são textos de comentário que se distinguem pela Atividade em que foram produzidos. Para o nosso trabalho, decidimos manter, na primeira fase, o COMJUR como um *subcorpus* do COMENTA, e, deste modo, poderemos analisar os dados separadamente ou em conjunto.

Para a segunda fase da análise, utilizámos como *corpus* o conjunto designado por COMENTA2 com 80 textos provenientes do G&T.Comenta e dois textos provenientes do COMJUR, totalizando 82 textos analisados:

	Tokens	Palavras	Número de segmentos de TD
COMENTA2	225.135	~180.558	188
COMJUR	6.884	~5.520	
CETEM2		~16.202	185
Total	232.009	202.280	373

Tabela 4: Corpus da segunda fase.

Na terceira etapa da análise usámos ainda uma seleção de textos do CETEM, a que demos o nome de CETEM2 e selecionámos 16 textos do COMENTA2. A seleção foi feita com base no número de segmentos de TD identificados e não no número de texto: deste modo, pretendeu-se equilibrar o número de TD entre o CETEM2 e o COMENTA2, tal como podemos observar na quarta coluna “Números de segmentos de TD” da Tabela 6.

A escolha dos textos para este *corpus* teve, também, como requisito incluir textos produzidos em contextos diferenciados (Jornalístico, Académico, Jurídico) e com temas variados (Sociedade, Ciência e Tecnologia, Religião, Economia, Direito e Literatura). Como os textos deste *corpus* pertencem todos ao género "Comentário", adicionámos um conjunto de textos do CETEM (CETEM2), que pertencem ao género "Notícia" para que pudéssemos obter um *corpus* de contraste, com textos de um género diferente. Este processo de seleção foi essencial para estratificar o *corpus* e, deste modo, garantir a variedade dos textos ao nível macro, isto é, variabilidade do tema e da atividade dos textos.

Este segundo conjunto de textos, como explicámos anteriormente, é constituído por segmentos de texto e não por textos completos, e necessitaram por isso, de serem selecionados de acordo com o número de segmentos para que houvesse um equilíbrio entre ambos. Assim, foi selecionado um conjunto de segmentos, que totalizam ~16202 palavras.

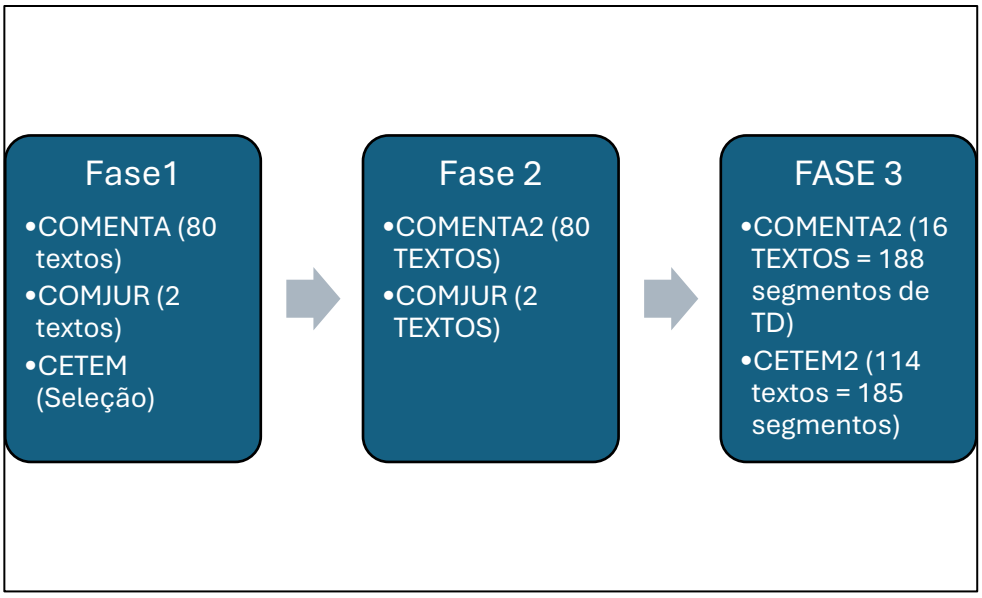


Figura 8: Síntese dos corpora usados em cada fase da investigação.

Como referimos anteriormente, as diferentes fases de análise implicam a extração de métricas diferentes. Para uma análise de fenómenos microlinguísticos é necessário equilibrar os *corpora* pelo número de palavras, uma vez que os textos do CETEM não correspondem a textos completos, mas a excertos. Na análise de fenómenos mesolinguísticos, como é o caso dos TD, o *corpus* COMENTA2 foi utilizado na íntegra,

enquanto na terceira fase, que analisou os fenómenos macro (contexto de produção dos textos), foi selecionada uma amostra.

Fazer a seleção de uma amostra tem alguns desafios, como notou McEnery (2001): determinar o tamanho da amostra implica decidir o tamanho do texto e o número de textos ideal a serem incluídos no *corpus*. Estes valores dependem da distribuição do fenómeno linguístico em análise dentro próprio *corpus*. Biber (1993) já tinha observado este problema, mostrando que as fórmulas estatísticas para determinar os tamanhos ideais das amostras e o número das amostras podem ser problemáticas. Os principais valores estatísticos para esta tarefa são o desvio padrão, calculado para cada característica individual, e o erro tolerável, que varia de acordo com a frequência geral de uma característica. De acordo com o autor, para calcular estes valores são necessárias várias etapas preparatórias que apresentam, elas próprias desvantagens.

Em primeiro lugar, impõe-se a necessidade de ter estimativas prévias do erro tolerável e do desvio padrão, o que por si pode ser problemático, porque muitas vezes não há, *a priori*, informações suficientes sobre a população antes de recolher uma amostra representativa. Em segundo lugar, o erro tolerável depende da precisão necessária para as análises que serão realizadas com base no *corpus*. No entanto, é difícil fornecer uma estimativa prévia da precisão necessária, especialmente em contextos onde a variação do fenómeno linguístico é alta. Em terceiro lugar, as equações podem não levar em conta a variabilidade entre as diferentes características linguísticas. Por exemplo, características raras podem ter desvios padrão muito diferentes, podendo resultar em tamanhos de amostra necessários que variam demasiado e que complicam ainda mais o *design* do *corpus*. As equações focam-se na representação de médias, mas não abordam adequadamente a diversidade linguística nos registos, o significa que, mesmo que uma amostra represente bem a média de certas características, pode não captar a variação e a diversidade que são cruciais para análises linguísticas mais abrangentes. Finalmente, a aplicação destas fórmulas estatísticas pode revelar-se um processo circular, onde investigações empíricas em *corpora* piloto enformam o processo de *design*, mas isso pode dificultar a definição de parâmetros iniciais porque estimar o desvio padrão de uma variável requer uma amostra representativa. Para atenuar estes problemas, Biber (1993) sugere um modelo de construção de *corpora* que seja dinâmico

e adaptativo, permitindo que a teoria e a prática se enformem mutuamente ao longo do processo.

O tamanho das amostras deve ser representativo das características em análise, tanto das que apresentam maior variação, como das que apresentam menor variação. E o número de textos deve refletir, dentro de cada género, o grau de variação da característica que ocorre dentro dos géneros em análise.

4.2.3. Formato dos Dados

O *corpus* CETEMPublico é disponibilizado em formato de texto simples (*.txt*), organizados verticalmente numa coluna, combinado num único documento. Como referimos na secção anterior, os artigos foram segmentados em extratos e são compostos por dois parágrafos, de tamanho variável, do artigo original (Rocha & Santos, 2000). Para este trabalho, construiu-se uma macro que reconstituiu os excertos utilizados no nosso *corpus*, para que obtivéssemos as sequências de texto originais. Este processo de reconstrução dos segmentos de texto do CETEM tem também como objetivo obter o contexto para as unidades linguísticas em análise.

O *corpus* G&T.Comenta tem uma origem mais heterogénea. Os textos foram recolhidos em suporte papel ou digital, que posteriormente foram digitalizados em formato PDF. Foi a partir do formato PDF, que recolhemos os textos e convertemos em ficheiros (*.txt*), através de um processo de OCR. Foi atribuído aos ficheiros um número de identificação que constituía o nome do ficheiro e, dentro de cada ficheiro, essa informação ficou disponível em formato XML (<ID> </ID>).

O XML (*eXtensible Markup Language*) é uma extensão restrita do SGML (Standard Generalized Markup Language) e foi desenvolvido principalmente para a troca de dados na internet. No contexto da linguística de *corpora* (Hardie, 2014), o XML serve como um sistema de marcação flexível e padronizado para estruturar, codificar e armazenar textos, permitindo que informações adicionais, como metadados e anotações analíticas, sejam integradas ao texto principal de forma sistemática. É, por isso, considerado uma das melhores opções para anotação em linguística de *corpora* por várias razões:

(i) O XML é uma versão mais simplificada do SGML, o que facilita o processamento e a implementação de anotação, não necessitando de um conhecimento técnico especializado para o executar.

(ii) A linguagem XML oferece uma estrutura padronizada que facilita a troca de dados entre sistemas diferentes e a integração de informações dentro do *corpus*, como por exemplo, *tags* da estrutura textual, metadados e anotações.

(iii) Como muitas ferramentas de análise de *corpora* são sensíveis ao XML, o uso deste formato facilita a análise automatizada dos textos, uma vez que a maior parte do *software* utilizado em linguística de *corpora* suporta XML, ou pelo menos parcialmente.

(iv) Comparado com o SGML, o XML omite muitas das características que tornam o SGML complexo de processar, o que o torna uma escolha mais prática e acessível, especialmente para projetos individuais ou de menor dimensão.

(v) Finalmente, a utilização do XML para a marcação de *corpora* oferece flexibilidade para ajustar a codificação de acordo com as necessidades específicas do estudo, sem a complexidade excessiva que os padrões mais pesados como o TEI (*Text Encoding Initiative*) exigem.

Os vários ficheiros *.txt* foram depois combinados num único documento com o qual foi constituído o *corpus* COMENTA2, utilizado na primeira fase.

Na segunda fase do nosso trabalho, utilizámos o *corpus* COMENTA2 para a análise dos tipos discursivos. Nesta etapa do trabalho, foi necessário proceder ao pré-processamento dos textos para que não houvesse falhas na leitura em XML. Assim:

- foram eliminados os parágrafos dos textos;
- foram normalizados todos os sinais de pontuação, em particular o tipo de aspas e os acentos;
- procedeu-se à substituição do sinal *&* e de parêntesis em cunha (<>) que estavam presentes no texto para que não interferissem na anotação XML.

Dada a dimensão do *corpus*, esta normalização foi feita com recurso a uma simples pesquisa e substituição. Depois do trabalho de pré-processamento, estabelecemos um esquema de anotação que permitisse recolher e tratar os dados

sobre os tipos discursivos. Para o estabelecimento de um esquema de anotação em XML, foi necessário fazer uma reflexão sobre os aspetos técnicos e teóricos que estavam subjacentes, e que colocavam alguns desafios que vamos detalhar nos parágrafos seguintes.

Na terceira etapa do trabalho, os segmentos de texto selecionados foram objeto de uma anotação, em XML, para organizar a informação relevante à criação do modelo. Foi criado um *element* com a identificação do texto <ID>, que contém o atributo[28] <TEMA>. Dentro deste *element* foram identificados e anotados manualmente os TD, onde cada *element* é identificado pelas siglas dos tipos discursivos (DT, DI, N, RI), numa estrutura hierárquica em que os tipos discursivos são nós de nível inferior, diretamente subordinados ao nó texto, e se configuram como unidades coordenadas entre si, no mesmo nível estrutural, como ilustrado no exemplo do texto 702:

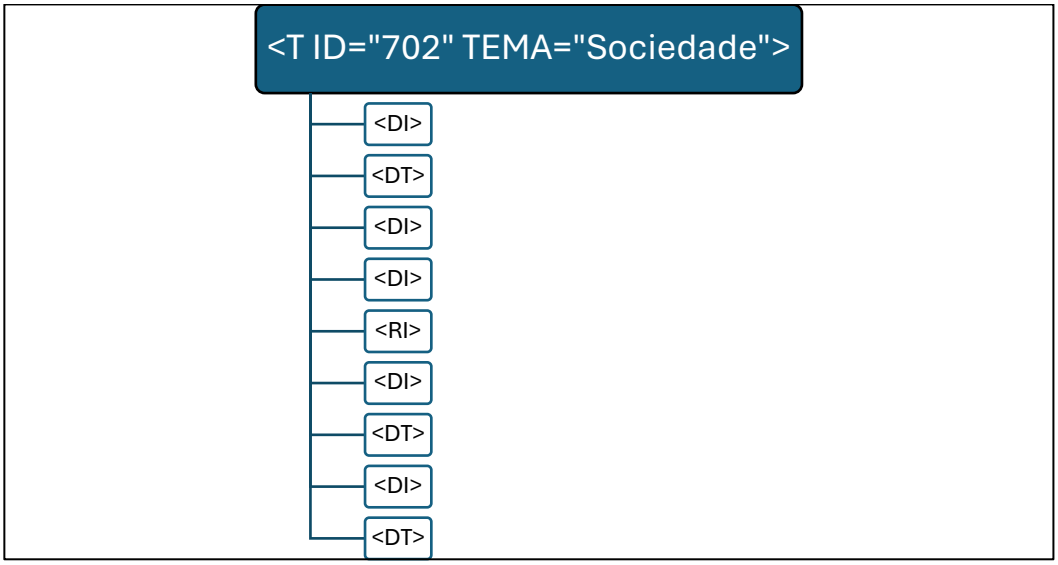


Figura 9: Esquema de anotação dos TD em XML.

O resultado desta anotação foi convertido para uma folha de cálculo, que nos permitiu obter o número dos tipos discursivos que ocorrem nos *corpora*, por género textual, e que se encontram descritos na Tabela 1:

[28] Em linguagem XML, o *element* designa uma unidade textual, visto como um elemento estrutural, enquanto os atributos servem para indicar informações que descrevem uma ocorrência específica de um elemento, mas que não são consideradas como parte do seu conteúdo (TEI Consortium, 2002: 22).

Género	Total	% no *corpus*
Comentário	188	50,40%
Notícia	185	49,60%
Total	373	100%

Tabela 5: Número e percentagem de tipos discursivos anotados por género textual.

De acordo com o que podemos observar na Tabela 1, aos 188 segmentos obtidos do *corpus* do COMENTA2, adicionámos 185 segmentos obtidos a partir do CETEM, num total de 373 segmentos. Relativamente à percentagem que cada um dos segmentos representa no *corpus* que vamos analisar, os segmentos pertencentes ao género Comentário representam 50,40%, enquanto os segmentos pertencentes ao género Notícia representam 49,60%.

Relativamente às variáveis para análise, o novo *corpus* apresenta quatro classes: Género, Atividade, TD e Tema. Em relação à Atividade, é importante sublinhar que todos os textos recolhidos do CETEM pertencem à atividade jornalística. Mantivemos as etiquetas do Tema atribuídas ao *corpus* CETEM (Rocha & Santos, 2000), que coincidiu em duas (Sociedade, Economia), e divergiu nas outras que foram adicionadas (Desporto, Política, Cultura, Opinião e Não Determinável [29](ND)) com os temas do COMENTA2.

A classificação do género (Notícia e Comentário) foi feita por nós. No caso do comentário, esta premissa foi baseada no facto de o *corpus* ter sido recolhido através de etiquetadas peritextuais ("comentário") ou referências à qualidade do autor enquanto "comentador" no peritexto ou no próprio texto. Para a notícia, assumimos que a etiqueta é interpretada num sentido mais lato, colocando o foco no meio de circulação (jornal diário).

[29] *Casos como Última Página e Destaque são classificados como ND (não determinável), visto que não considerámos obrigatório atribuir classificações.* (Rocha & Santos, 2000: 5)

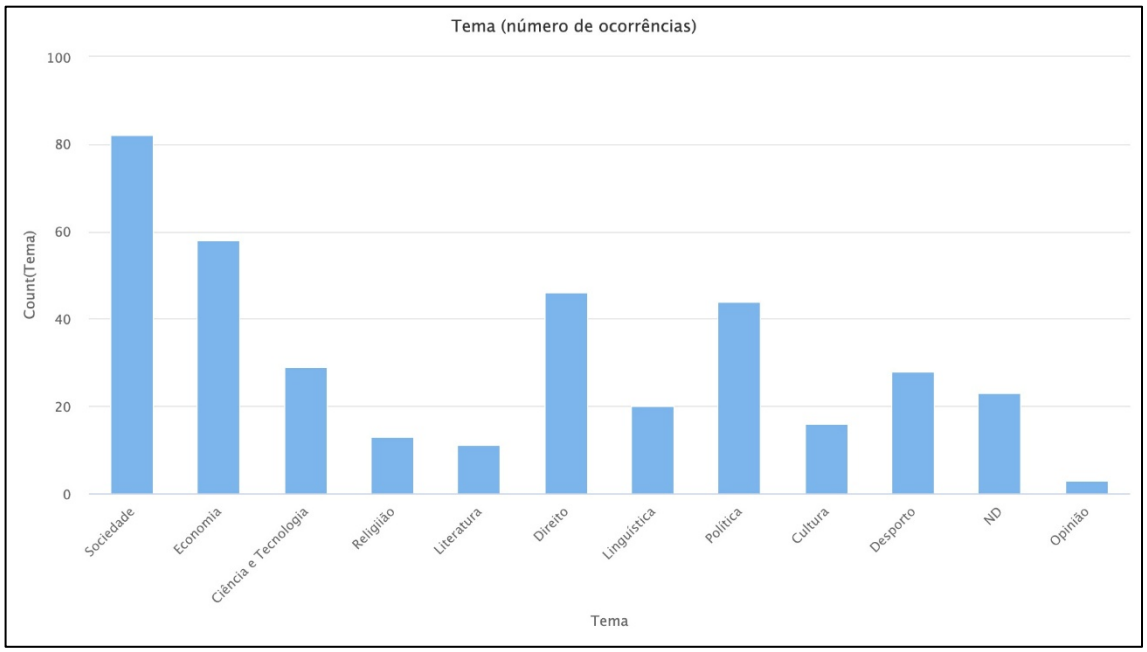


Gráfico 1: Número de documentos por Tema.

Como podemos observar pelo Gráfico 1, obtivemos 12 temas diferentes para este novo *corpus*: existe uma representação maior dos temas que coincidiram, Sociedade e Economia, e uma menor representação de temas que não coincidiram em ambos os *corpora*. Poderíamos discutir alguns aspetos desta classificação, como por exemplo os critérios de atribuição dos temas e se a etiqueta "Opinião" pode ser considerado um tema ou reflete mais uma prática textual, com características próprias. Para o nosso trabalho, vamos assumir as escolhas feitas, nesta classificação, pelos autores do *corpus* CetemPublico (Rocha & Santos, 2000).

As classes estabelecidas, que constituirão doravante as variáveis, estão descritas na tabela seguinte:

ID SEQ	Numeração sequencial que identifica cada um dos segmentos de texto. Um para um.
ID	Número de identificação do texto no *corpus*. Um para muitos.
Tema	Tema geral do texto.
Atividade	Atividade em que foi produzido o texto.
TD	Tipo discursivo atribuído ao segmento de texto
Género	Género atribuído ao texto.

Tabela 6: Atributos para análise.

As classes ID SEQ e ID têm apenas uma função identificativa e não serão usadas na construção do modelo. A classe ID SEQ é um número sequencial atribuído a cada um dos segmentos de TD identificados, enquanto a classe ID identifica cada um dos textos. Os dados quantitativos, de cada umas das classes, podem ser consultados no Anexo 6.

4.2.3.1. Anotação em XML no âmbito da classificação automática de texto

O uso de *corpus* anotados para a análise linguística tem conhecido um crescimento significativo nas últimas décadas, à medida que as capacidades técnicas têm disponibilizado novas ferramentas de análise à linguística. Se inicialmente, a anotação de *corpora* era um projeto de grande envergadura que exigia grandes equipas e financiamento, atualmente, com a democratização do acesso a ferramentas informáticas, a constituição e uso de *corpora* anotados para estudos de menor dimensão e mais específicos têm-se popularizado.

Os primeiros trabalhos de estandardização de anotação eram, precisamente, orientados para a anotação de grandes *corpora*, como por exemplo *Text Encoding Iniciative* (TEI) e o *Expert Advisory Group on Language Engineering Standards* (Eagles). No entanto, como referido em trabalhos anteriores (Hardie, 2014), o formato TEI foi criado e desenvolvido num momento em que a criação de *corpora* era uma tarefa desenvolvida em projetos de investigação com equipas numerosas e, por isso, a anotação foi desenvolvida para ser profunda e para que tivesse o maior número de usos

possíveis. Um dos exemplos desta prática é o British National *Corpus* (BNC) que foi desenvolvido sob a alçada de um consórcio composto por instituições públicas e privadas, e cujas equipas eram compostas por linguistas de várias áreas, bem como programadores informáticos. O BNC tornou-se, por isso, um *corpus* de referência para a investigação em linguística do inglês (britânico). Para projetos de anotação mais pequenos, ou mesmo individuais, têm surgido algumas propostas que, não abandonando o TEI ou o Eagles, adaptam e simplificam as normas. Entre os exemplos de propostas que se têm surgido, destacamos (Blache et al., 2008) pela tentativa de anotar *corpora* multimodais, ao mesmo tempo que procura lidar com questões mais teóricas, como, por exemplo, a variação da terminologia aplicada, de acordo com o quadro teórico usado. O outro exemplo de proposta que destacamos é (Goecke, Liingen, Metzing, & Stiihrenherg, 2010) que apresenta uma proposta para a representação da informação. Este trabalho é particularmente importante porque introduz a noção de representação da informação por níveis, considerando que a natureza da informação obriga a ferramentas distintas. Assim, e de uma forma simplificada, os autores designam por nível como o processo conceptual que parte de um conceito teórico (sintaxe, semântica, entre outros), enquanto a camada designa a execução técnica da anotação e que, por sua vez, parte do sistema de anotação escolhido. A Figura 10 mostra como os níveis e as camadas relacionam o nível conceptual com a execução técnica.

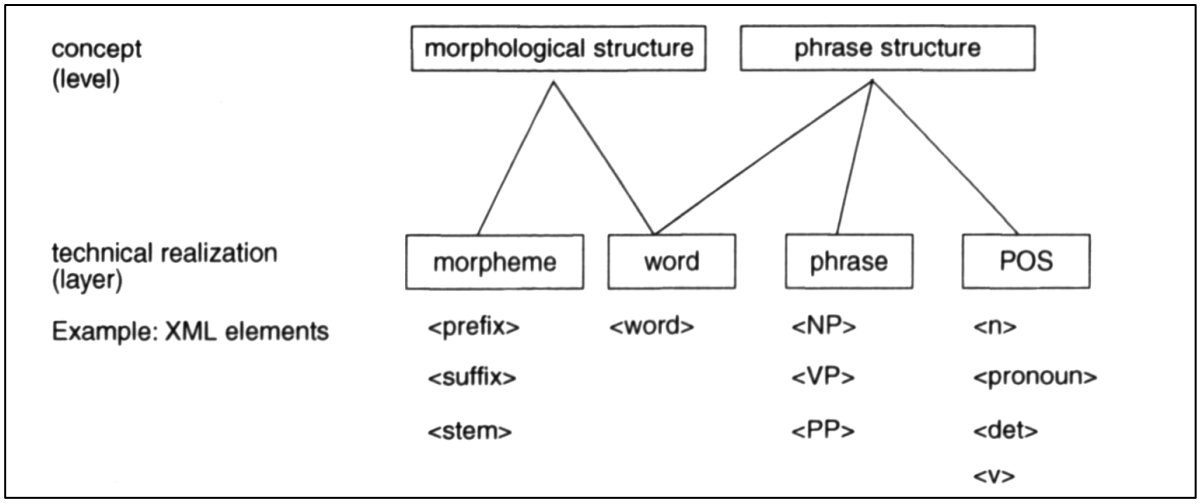


Figura 10: extraído de Goecke, Liingen, Metzing, & Stiihrenherg (2010: 3)

De acordo com os autores, e explicando o esquema da Figura 10, existem várias relações possíveis entre níveis e camadas:

(i) Relação um para um: Um nível de anotação corresponde diretamente a uma camada de anotação. Significa que para cada nível conceptual, há uma camada técnica que o representa, e uma camada pode ser removida ou trocada sem afetar as outras.

(ii) Relação um para muitos: Um nível de anotação é distribuído por várias camadas. Por exemplo, para o nível de partes do discurso (POS), podemos ter diferentes camadas para cada classe de palavras (substantivos, verbos, etc.).

(iii) Relação vários para muitos: É criado quando vários níveis conceptuais são integrados numa única camada de anotação. Ocorre quando diferentes tipos de informação (como sintaxe e morfologia) são combinados numa única representação técnica (como um único documento XML).

(iv) Relação vários para muitos: É a relação mais complexa e implica a divisão e mistura de níveis conceptuais. Embora não seja comum, pode ocorrer quando diferentes tipos de informação são combinados e distribuídos em múltiplas camadas.

Sendo o XML uma linguagem de marcação, esta é formatada com um conjunto de restrições, ou sintaxe, das quais consideramos a mais relevante o facto da anotação em XML obrigar a que os elementos sejam corretamente estruturados na mesma camada. Esta restrição coloca alguns obstáculos uma vez que nem sempre a informação anotada, sobretudo de for de caráter semântico, é contínua. Em (Magalhães & Gonçalves, 2021) foi feita uma reflexão sobre esta questão a partir da anotação da *deixis*. Os problemas que são elencados para a *deixis* centram-se, principalmente, no facto de a *deixis* ser um operador semântico, que depende de um conjunto de operações, de natureza abstrata, com fatores *intra* e extralinguísticos. Na secção seguinte, vamos elencar e analisar alguns desafios que a anotação dos tipos discursivos coloca.

4.1.3.2. Anotação em XML de Tipos Discursivos

No nosso trabalho, e em especial nesta etapa, a questão coloca-se apenas em parte: como referimos anteriormente, os tipos discursivos (TD) organizam-se em torno

de dois eixos que correspondem a diferentes planos de enunciação. Por um lado, temos a relação com o agente (implicação ou autonomia) e, por outro, temos a organização temporal relativamente ao ato enunciativo (conjunção ou disjunção). Como está resumido na Tabela 3 do capítulo anterior, a *deixis* é um dos elementos através dos quais podemos aceder aos mundos discursivos / tipos discursivos, embora não seja o único elemento, uma vez que, como referimos anteriormente, para identificar um TD não é suficiente analisar a presença ou ausência das unidades linguísticas, sendo necessário também observar as várias unidades presentes no segmento textual e como interagem. Sendo os TD segmentos dotados de regularidades linguísticas, é necessário que a anotação destes segmentos seja o mais abrangente possível, e que possa incluir segmentos de textos de tamanho variável que pode incluir partes de frases ou frases completas, mas também um ou mais parágrafos. Vejamos, como exemplo, a distribuição dos tipos discursivos do **Texto 2** (Anexo 1):

Chegámos à CES 2015, Consumer Electronics Show, de Las Vegas, com alguma expectativa sobre os televisores quantum dot LED. O entusiasmo esmoreceu ao confirmarmos que o dito novo tipo de iluminação de ecrã já tinha sido usado nalguns modelos da série Triluminos da Sony, em 2013. Tínhamos testado alguns e confrontámos a teoria com os resultados obtidos na altura, em laboratório.	**RI**
Cores mais nítidas, naturais e com uma maior gama: a promessa dos microscópicos quantum dots, com 2 a 10 nanómetros, baseia-se na capacidade de reproduzir as cores com maior precisão. As suas minúsculas dimensões permitem comprimentos de onda menores, traduzindo-se em cores mais precisas. Se expostos à luz azul emitida pelos LED, os quantum dots convertem alguma dessa luz, na origem, para verde e vermelho. O resultado na imagem abrange as três cores primárias.	**DT**
Quantum dot LED divide opiniões Escrutinámos os resultados dos modelos da Sony com quantum dot LED. Ao contrário dos televisores LCD LED, nos quantum dot, a perceção das cores divide um pouco o painel que visualiza as imagens dos televisores no nosso teste. Alguns dão boa nota às cores, mas outros relatam desvios nos tons. Sentados frente a um ecrã, quantum dot para ver uma imagem só com as cores mais puras, ficaríamos impressionados com a tecnologia. Mas nas imagens de televisão, como na vida, não vemos só cores puras.	**RI**
Também há misturas de cores ou cores pastel, como nos tons de pele. É aí que a diferença de opiniões começa.	**DT**

Tabela 1: Distribuição dos tipos discursivos

Este texto, em que o enunciador é um jornalista especialista em tecnologia, recorre a segmentos com características do Relato Interativo (RI) e do Discurso Teórico (DT). Como podemos observar, os segmentos têm um tamanho variável, mas são constituídos por blocos mais ou menos independentes[30].

A articulação dos TD é um aspeto a ter em conta para este trabalho, uma vez que a anotação em XML estabelece uma hierarquia entre os elementos, que é recorrente na anotação de textos. Várias propostas (Cristea & Butnariu, 2009; Goecke et al., 2010) têm sido feitas para resolver esta questão e, para o nosso processo de anotação, decidimos que o processo mais adequado é o da fragmentação (Goecke et al., 2010: 5), em que uma sequência de texto incluída num elemento que de outra forma seria afetado por uma sobreposição é dividida em várias sequências de texto. Deste modo, a anotação assume um esquema mais simples, em que cada tipo discursivo é um nó[31] irmão. Na Figura 11, podemos observar um exemplo de um TD, no caso DI, que surge integrado numa frase ou parágrafo do tipo N, marcada por uma fronteira gráfica (vírgula).

```
            <N>
342             Para Antero de Quental (1842–1891), o género epistolar foi um lugar
343 geográfico soberano (invadido por tantos outros) de uma biografia incumprida,</N>
            <DI>
344             o que, de imediato, levanta a questão de saber em que consiste essa
345 unidade descontínua daquilo que designamos por obra. Será que tudo o que um autor
    escreveu o é, correspondência incluída? </DI>
            <DT>
346             O estabelecer desses parâmetros dir-se-ia também tarefa de quem
347 investiga um espólio, tentando salvaguardá-lo de um eventual desaparecimento,
    questão que, desde Mallarmé, tem vindo a ser pensada e repensada e que Foucault
    abordou tendo em conta o manancial de lacunas e fissuras, por onde se perscrutam
    espaços, que toda a escrita possui.</DT>
            <DI>
348             Cinjamo-nos, então, ao epistolário de Antero, que agora surge na
349 sua quarta edição*, exaustiva e rigorosamente tratado, desde os anos 80, pela
    investigadora Ana Maria Almeida Martins, autora da fotobiografia e fiel
    anteriana. </DI>
```

Figura 11: Exemplo de anotação XML.

Os exemplos que aqui convocados mostram que a identificação dos TD implica uma análise qualitativa dos textos, que passa pela anotação manual, para obter dados que possam ser analisados estatisticamente. Esta análise qualitativa obriga, em algumas situações, a tomadas de decisão por parte do anotador que podem levar à

[30] Em relação à forma como os TD se articulam entre si (fusão ou encaixe).

[31] Um nó (*node*) refere-se a qualquer elemento, atributo, texto ou comentário no documento XML.

perda de alguma informação semântica, em detrimento dos dados estatísticos. Como referem (McEnery & Wilson, 2001):

> *To ensure that certain statistical significance tests (...) provide reliable results, it is essential that specific minimum frequencies are obtained, and this can mean that fine distinctions have to be deliberately blurred to ensure that statistical significances can be computed, with a resulting loss of data richness.*

No caso dos tipos discursivos, esta situação coloca-se principalmente quando ocorre uma articulação em fusão, exemplificada na Figura 11, onde podemos observar a integração de um discurso interativo com um discurso teórico, própria de uma função didática ou da divulgação científica, em que a informação científica, de caráter autónomo surge intercalada com texto implicado, que procura apelar ao seu interlocutor. Esta fusão é particularmente evidente no parágrafo que constitui a linha 153 da Figura 12, em que a apresentação da informação, própria da autonomia do discurso teórico, apela simultaneamente ao recetor, convocando caraterísticas da implicação. Este é um exemplo da dificuldade da anotação dos TD, sobretudo quando a articulação ocorre em fusão. Menos problemático para a anotação, o encaixe de vários tipos discursivos surge delimitado por fronteiras mais ou menos explícitas, geralmente pelo uso de marcas gráficas que separam frases ou parágrafos, que segmentam o texto, como exemplificamos com a representação do **Texto 2** em XML:

```
    <T ID="718">
        <RI>
            Comentário do especialista. CES Las Vegas. Chegámos à CES 2015,
Consumer Electronics Show, de Las Vegas, com alguma expectativa sobre os
televisores quantum dot LED. O entusiasmo esmoreceu ao confirmarmos que o dito
novo tipo de iluminação de ecrã já tinha sido usado nalguns modelos da série
Triluminos da Sony, em 2013. Tínhamos testado alguns e confrontámos a teoria com
os resultados obtidos na altura, em laboratório. </RI>
        <DT>
            Cores mais nítidas, naturais e com uma maior gama: a promessa dos
microscópicos quantum dots, com 2 a 10 nanómetros, baseia-se na capacidade de
reproduzir as cores com maior precisão. As suas minúsculas dimensões permitem
comprimentos de onda menores, traduzindo-se em cores mais precisas. Se expostos à
luz azul emitida pelos LED, os quantum dots convertem alguma dessa luz, na
origem, para verde e vermelho. O resultado na imagem abrange as três cores
primárias.</DT>
        <RI>
            Quantum dot LED divide opiniões. Escrutinámos os resultados dos
modelos da Sony com quantum dot LED. Ao contrário dos televisores LCD LED, nos
quantum dot, a perceção das cores divide um pouco o painel que visualiza as
imagens dos televisores no nosso teste. Alguns dão boa nota às cores, mas outros
relatam desvios nos tons. Sentados frente a um ecrã quantum dot para ver uma
imagem só com as cores mais puras, ficaríamos impressionados com a tecnologia.
Mas nas imagens de televisão, como na vida, não vemos só cores puras. </RI>
        <DT>
            "Também há misturas de cores ou cores pastel, como nos tons de
pele. é aí que a diferença de opiniões começa. António Alves. PRODUTOS E
SERVIÇOS. "Os poucos modelos quantum dot LED analisados dividiram um pouco as
opiniões no painel, o que não Sucedeu nos LCD LED".</DT>

    </T>
```

Figura 12: Anotação em formato XML

Neste exemplo podemos observar que os segmentos dos TD estão, de uma maneira geral, delimitados por marcas gráficas, como o ponto final e o parágrafo, que assinalam o início e o fim dos segmentos.

4.2. Métodos de Análise

4.2.1. Ferramentas Utilizadas

Na primeira fase, o *corpus* foi analisado através da ferramenta *SketchEngine* (Kilgarriff, Tugwell, Rychly, & Smrz, 2004); (Kuhn, 2019). O software *SketchEngine* é uma ferramenta paga, acedida por nós através de uma licença académica, e é uma das ferramentas de *corpus* mais robusta disponível até ao momento. Além da gestão de *corpora online*, o *SketchEngine* é também uma ferramenta de *Corpus Query Tool* (CQL), com uma extensa lista de funcionalidades. Não estando no propósito deste trabalho demonstrar todas as funcionalidades do *SketchEngine*, vamo-nos debruçar sobre as ferramentas de pesquisa mais utilizadas no nosso trabalho.

Depois de o *corpus* ser carregado em formato *.txt*, procedemos à sua gestão, subdivindo-o em três *subcorpus*: (i) CETEM, (ii) COMENTA2 e (ii) COMJUR.

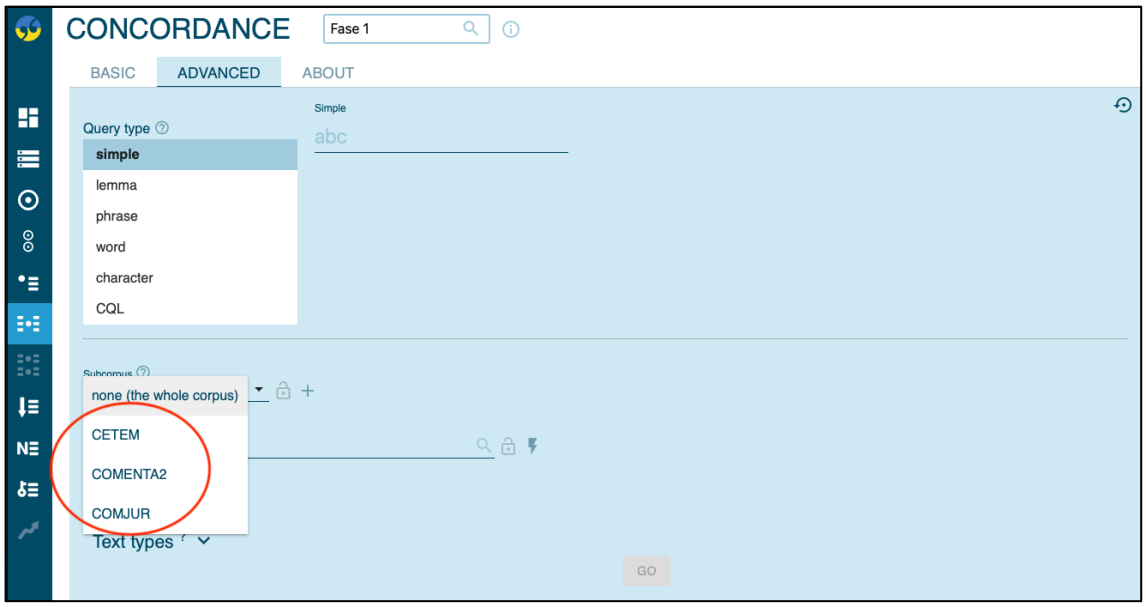


Figura 13: Painel de pesquisa do Sketch Engine.

Na Figura 13 podemos observar o menu de pesquisa avançado do *Sketch Engine* onde podemos selecionar um dos três *subcorpora* ou o *corpus* completo. O *Sketch Engine* permite também vários tipos de consulta, com diversos graus de resultados, que podem ir da pesquisa simples de uma palavra, *lemma* ou frase, ou executar pesquisas mais complexa, através do CQL, filtrando os atributos desejados, através da anotação do *corpus*. O *tagset* usado pelo Sketch Engine para a anotação encontra-se disponível no Anexo 11. Para obter os dados que nos propomos analisar, são necessárias várias pesquisas de expressões complexas. Observemos alguns exemplos: Para obter, por exemplo, o número de advérbios de lugar[32] com possível interpretação deítica, podemos executar uma pesquisa simples ([word="aqui | aí | ali | cá | lá"]) e o programa retorna uma concordância com as ocorrências.

[32] Para o exemplo: *aqui, aí, ali*.

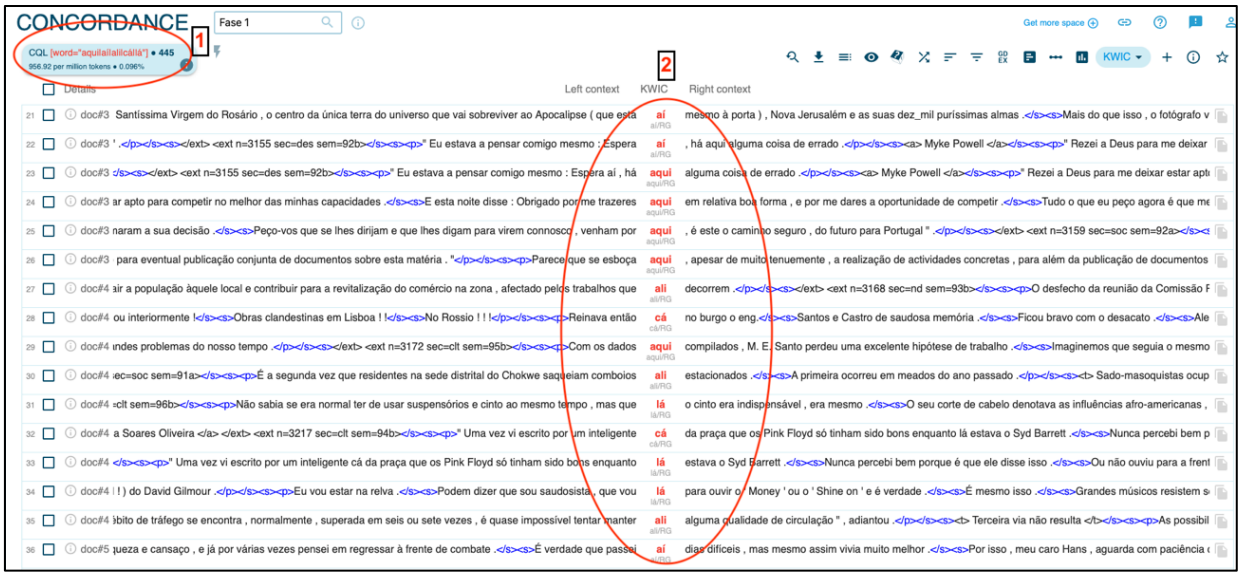


Figura 14: Resultados de uma concordância.

O quadro 1 tem um caráter informativo e permite-nos obter alguns dados estatísticos da pesquisa efetuada. Neste caso obtemos um total de 445 ocorrências, que representam 0.09569% do *corpus* total. É partindo destes números que vamos obter os dados estatísticos da primeira fase. No menu 2, obtemos a concordância com a lista de palavras que pesquisámos. Para fazer, por exemplo, um tempo verbal composto, como o pretérito perfeito composto, podemos executar o CQL **[lemma="ter"& tag="VMIP.*"]+[tag="V.*"]**, em que:

- **[lemma="ter"** - *para obter o verbo auxiliar;*
- **& tag="VMIP.*"]** - *para que o auxiliar "ter" esteja apenas no presente do indicativo;*
- **+[tag="V.*"]** - *para que seja seguido de uma qualquer palavra que tenha a etiqueta "verbo".*

A concordância resultante pode ser vista na Figura 15. A lista de todas as expressões CQL usadas nesta fase do trabalho encontra-se listada detalhadamente no Anexo 3.

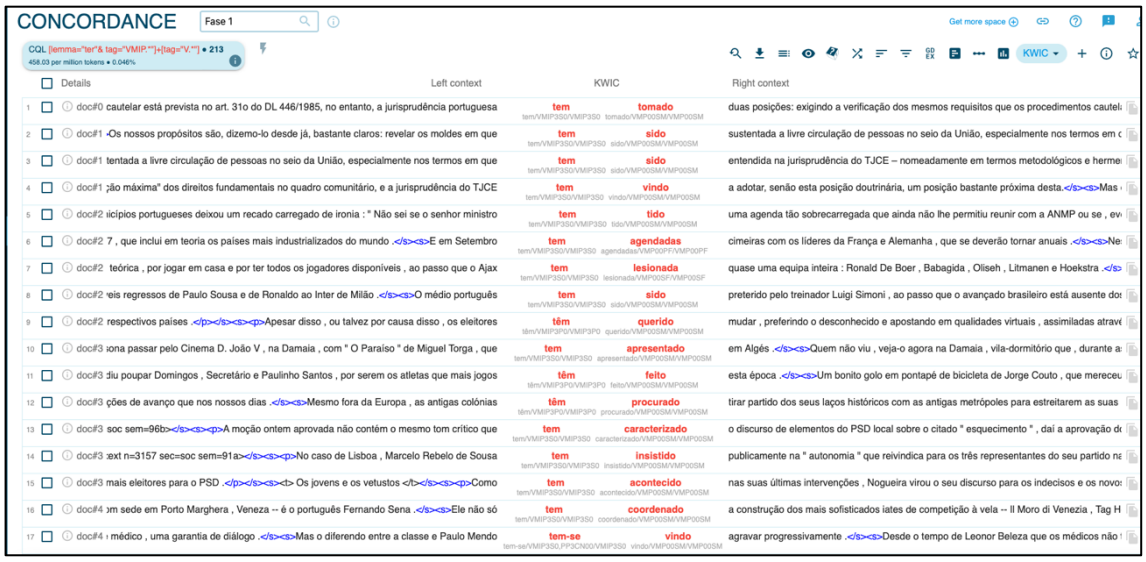


Figura 15: concordância do pretérito perfeito composto.

Na segunda etapa da nossa investigação, que incide sobre os tipos discursivos, optámos por desenvolver a análise estatística em *Python*. Para isso, foram desenvolvidos *scripts* que se adaptassem ao tipo de análise que íamos desenvolver. Como referimos na secção **4.1.3.2**, foi necessário fazer uma anotação manual dos tipos discursivos, em formato XML, e para fazer a análise estatística deste tipo de anotação, o *Python* revelou-se a opção mais rápida e flexível (Bird, Klein, & Loper, 2009; Perkins, 2014).

Na terceira fase do nosso trabalho, que implica uma análise multivariada (Biber, 1995a, 2004), utilizámos o programa de *data mining RapidMiner Studio* (versão *Educational* 10.3), versão gratuita, renovável, que permite mais de 10 mil exemplos, ao contrário da versão *Trial*, também gratuita, mas com limitações. Existem vários programas de *data mining*, tanto pagos como gratuitos, e o *RapidMiner* mostra algumas vantagens, dos quais destacamos:

- O *RapidMiner* é constituído por módulos expansíveis através de bibliotecas disponibilizadas gratuitamente (como por exemplo bibliotecas específicas para *text mining*).
- Permite otimizar os processos através da adição de códigos em *Python* e *R*: é aberto e extensível através da integração de bibliotecas.

- O *RapidMiner* permite ligar / utilizar qualquer tipo de base de dados (XML, Excel, entre outras), e permite recuperar informações de base de dados sem o uso de linguagens *SQL* complexas.
- Os operadores utilizados não funcionam em *blackbox*, deste modo, cada etapa da análise é documentada, de forma completamente transparente.
- A interface do programa é intuitiva, embora seja necessário algum conhecimento básico de *text e data mining*.

O *RapidMiner* surgiu, por estes motivos, como a opção mais viável para os objetivos propostos.

O primeiro passo para a análise com *RapidMiner* passa pela ligação à base de dados que, no nosso caso, foi convertida numa folha de cálculo (.XLS). No ecrã inicial do programa, que corresponde à janela de *Design*, partindo de um projeto novo e não utilizando os *Templates* disponíveis, encontramos três áreas fundamentais: (1) o menu dos operadores, (2) a janela de processo e (3) o menu de parâmetros:

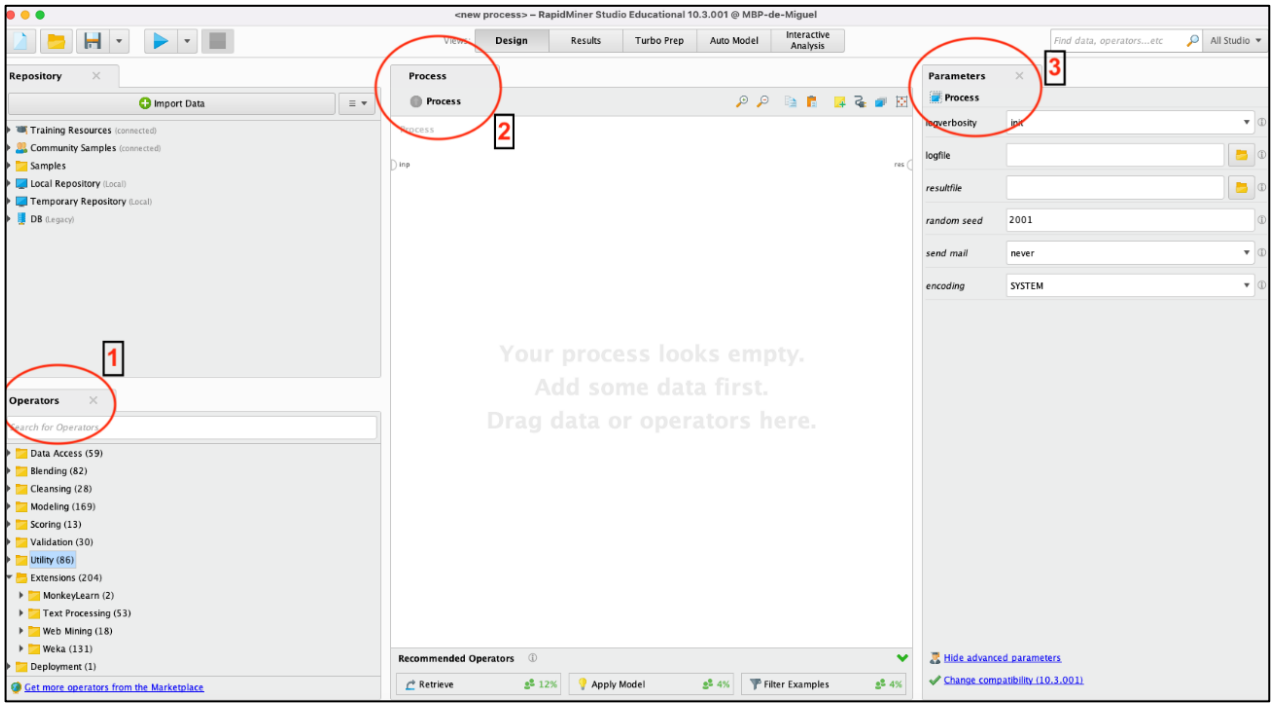


Figura 16: Menu principal do programa RapidMiner.

O menu de operadores é uma biblioteca organizada de ferramentas que podem ser usadas para realizar tarefas específicas em análises de dados. Cada operador representa uma função, como carregar os dados, transformar as variáveis, treinar os

modelos de *machine learning*, ou avaliar resultados. É no menu dos operadores que adicionamos as funções necessárias ao processamento dos dados, num sistema de *pipeline*[33] que podemos visualizar no menu 2. No menu 3 onde podemos ajustar os parâmetros de funcionamento de cada operador, configurando os detalhes específicos de cada um dos operadores. Enquanto no menu 1 definimos o que será feito, no menu 3 definimos como será feito.

Para a segunda fase utilizámos um operador para aceder e ler os dados da folha de cálculo e um segundo operador para filtrar dados. No nosso caso, usámos um filtro para excluir os tipos discursivos que ocorrem em citações. Esta opção foi tomada porque a citação, embora seja um processo discursivo com diversos valores, que pode ser considerado um marcador de género, ocorre, no *corpus* em análise, como uma legitimação do conteúdo temático (no caso da atividade académica e jurídica) e na indicação da fonte ou introdução de novos temas (Miranda, 2010a: 273-274). Mas como os valores e os tipos discursivos que ocorrem neste contexto são muito variáveis, decidimos manter dois cenários: o primeiro cenário exclui os tipos discursivos em contexto de citação, e tem como objetivo quantificar os tipos discursivos que surgem apenas no texto comentário, sem contabilizar os tipos discursivos a que pertencem as citações. O segundo cenário ocorre durante a análise das sequências dos tipos discursivos. A decisão de efetuar a análise neste cenário tem como objetivo perceber que tipos discursivos se encontram contíguos às citações, e se existe algum padrão relevante desta estrutura no comentário.

Depois de executar os processos, passamos para a janela dos Resultados, que mostra várias hipóteses de visualização dos dados obtidos pelo processo executado, e que podemos observar na Figura 17. No menu 1 podemos aceder aos resultados numa tabela. No menu 2 podemos observar uma pré-visualização e síntese estatística dos atributos analisados, e no menu 3 acedemos à visualização dos dados através de várias

[33] Entendemos por processo *pipeline* uma sequência de etapas ou processos interligados, onde a saída de uma etapa é a entrada da próxima. Este tipo de processo implica dividir o processo de análise em várias etapas distintas e encadear essas etapas numa sequência lógica, que permite uma abordagem modular e escalável para lidar com grandes volumes de dados e complexidade analítica.

opções de gráficos, e onde podemos integrar e visualizar os diferentes atributos, e observar como se relacionam entre si. É, talvez, uma das ferramentas mais úteis ao nosso trabalho, uma vez que de uma forma intuitiva permite alterar e observar a interação dos vários atributos de uma forma acessível.

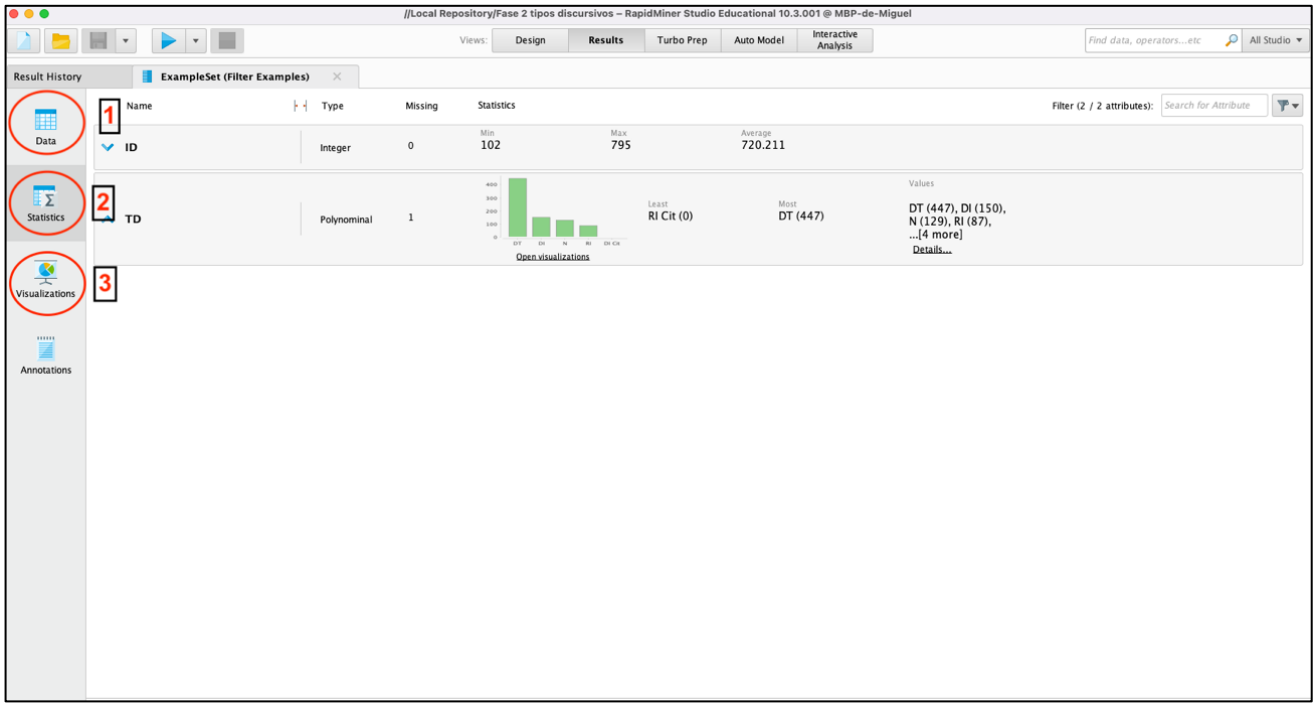


Figura 17: Ecrã dos resultados.

4.3. Procedimentos Analíticos

Na primeira fase deste trabalho, procedeu-se à criação de uma lista de "etiquetas" que pudessem descrever as unidades linguísticas do *corpus*. O objetivo desta fase do trabalho foi fazer uma representação dessas unidades linguísticas e gramaticais do comentário e perceber se existe alguma especificidade das categorias analisadas, através de um processo baseado, mas não idêntico, à estilometria (*stylometrics*) (Aaronson, 2001). Para coordenarmos esta fase com os objetivos do nosso trabalho, selecionámos um conjunto de características que pudessem, posteriormente, ser comparadas. Assim, procurámos descrever as unidades linguísticas que podem oferecer pistas para a descrição dos tipos discursivos, como descrevemos na primeira parte (**1.2.6.**) do nosso trabalho, nomeadamente a oposição entre implicado e autónomo no que toca à organização atorial e à conjunção ou disjunção respeitante a organização temporal. Com a definição destas variáveis, procurámos definir os seguintes aspetos: a distribuição lexical, a caracterização morfossintática dos verbos (os valores de pessoa-

número, e tempo-modo-aspeto), a caracterização pronominal (funções sintáticas) e os advérbios de tempo e lugar com possível interpretação deítica. A lista completa dos elementos pesquisados está disponível no Anexo 3.

Tentámos, nesta fase do trabalho, conjugar uma caracterização linguística mais genérica com uma análise de unidades linguísticas que nos pudessem fornecer elementos mais específicos do *corpus*. Bronckart (1988) sublinha a importância da função distintiva das unidades linguísticas em análise, que é validável pelo juízo dos locutores. Deste modo, as unidades linguísticas são analisadas primeiramente pela sua distribuição (critério paradigmático) e pela interdependência (critério sintagmático). Esta análise é feita adotando o texto enquanto estrutura de análise, dentro do qual podemos classificar as unidades. O segundo aspeto destacado pelo autor é a capacidade de uma mesma unidade poder veicular várias funções, e as mesmas sequências de unidades poderem traduzir (ou representar) funções diferentes. Também nesta questão, seguiremos a metodologia proposta de fazer, primeiramente, um levantamento das unidades (análise quantitativa) e proceder à análise das funções que assumem (análise qualitativa).

Depois de obtidos os dados quantitativos, foram feitas diversas análises. Em primeiro lugar foram construídas tabelas com os dados absolutos, que são mostrados através de gráficos de barras. Posteriormente, os dados foram alvo de um teste ANOVA[34] para apurar a sua significância estatística, utilizando, para isso um código *Python*[35].

[34] O teste estatístico ANOVA (Análise de Variância) é uma técnica utilizada para comparar as médias de três ou mais grupos e determinar se há diferenças estatisticamente significativas entre essas médias. Para o nosso estudo foi utilizado a ANOVA de uma via (One-Way ANOVA), utilizada quando há uma única variável independente (fator) com três ou mais níveis (grupos) e se deseja comparar as médias desses grupos. Se o valor de *F* for alto, indica que a variabilidade entre as médias dos grupos é maior do que a variabilidade dentro dos grupos, indicando uma possível diferença significativa entre as médias dos grupos. Se o valor de *F* for baixo, então a variabilidade entre as médias dos grupos não é significativamente maior do que a variabilidade dentro dos grupos. Para interpretar o valor de significância (valor *P*), se o Valor *P* for pequeno (geralmente < 0.05), então há uma diferença significativa entre as médias dos grupos. Se, pelo contrário, o valor *P* for grande (geralmente > 0.05), não há uma diferença significativa entre as médias dos grupos. (James, Witten, Hastie, & Tibshirani, 2023: 118-119).

[35] O teste ANOVA foi executado através de um *script* em *Python* com a biblioteca do scipy.stats.

A escolha do teste ANOVA como método estatístico para análise dos dados obtidos fundamenta-se (Gelman & Hill, 2007) na sua capacidade de avaliar diferenças significativas entre as médias de múltiplos grupos de unidades linguísticas extraídas do *corpus*. A ANOVA permite examinar se diferentes condições experimentais, neste caso, as atividades em que são produzidos os textos, influenciam de maneira estatisticamente significativa as unidades linguísticas analisadas. Além de permitir a comparação entre múltiplos grupos de unidades linguísticas, a ANOVA oferece uma abordagem robusta para garantir uma análise estatística rigorosa e a interpretação adequada das relações entre as características linguísticas das unidades no contexto. As duas principais desvantagens (Field, 2017; Gelman & Hill, 2007) deste teste são:

(i) a ANOVA assume que os dados de cada grupo estão normalmente distribuídos e que há homogeneidade de variâncias entre os grupos;

(ii) não especifica quais os grupos que apresentam diferenças. Significa, portanto, que se a ANOVA for significativa, são necessários testes complementares para identificar os grupos que diferem.

Num segundo momento, executámos outro teste que pudesse revelar algum tipo de padrão entre os *subcorpora*. Como não era possível, nesta fase do trabalho, perceber se existiam características que pudessem ser agrupadas (*clusters*), o primeiro teste que aplicámos foi o *hierarchical clustering*. O *hierarchical clustering* é uma técnica de *clustering* que permite vários níveis de agrupamento dos dados: os dados são divididos em vários grupos que se sobrepõem, numa espécie de árvore. A vantagem do *hierarchical clustering* é que oferece uma estrutura mais flexível e detalhada, permitindo ver os dados em diferentes níveis de granularidade (Tan, Steinbach, & Kumar, 2018).

As técnicas de *clustering* são usadas para dividir dados em grupos, criando etiquetas a partir de cada *cluster* (Feldman & Sanger, 2007) Neste caso, usámos o *hierarchical clustering*, utilizando a distância euclidiana[36] e o método de ligação de

[36] A distância euclidiana é a distância direta entre dois pontos num espaço multidimensional (Feldman & Sanger, 2007: 85).

Ward[37], para agrupar as variáveis num dendrograma. Este método é eficaz para identificar estruturas hierárquicas e criar *clusters* compactos com variância mínima interna (Tan et al., 2018). Entre as suas vantagens estão a flexibilidade no uso de medidas de similaridade e a clareza na visualização hierárquica dos dados. No entanto, existem algumas desvantagens, como a alta complexidade computacional e a sensibilidade a *outliers*, o que pode afetar a qualidade dos *clusters* formados. A técnica de *hierarchical clustering* é especialmente útil na análise de texto uma vez que permite visualizar a relação entre vários *corpora* (McEnery, 2001: 91). Para executar este teste, foi feito um *script* em *Python* que gerou um dendrograma, mostrando a relação hierárquica e a relação de semelhança entre os diferentes *subcorpora*.

Na segunda fase do nosso trabalho analisamos os tipos discursivos presentes no COMENTA2. Para fazer este tipo de análise, dividimos o processo em duas etapas: a primeira etapa foca-se na análise das frequências dos TD e a segunda centra-se nas sequências dos TD. Esta etapa foi realizada manualmente, através da análise dos textos.

Num primeiro momento vamos agrupar e contabilizar os tipos discursivos (DI, DT, RI, N) por texto (ID) que será expresso por um gráfico[38] com a distribuição dos TD, e numa tabela de frequência. Vamos também, partindo destes dados fazer uma matriz de correlação: A matriz de correlação, no contexto da análise multivariada, é usada para examinar as relações estatísticas entre várias variáveis, revelando como as diferentes variáveis se relacionam entre si nas suas distribuições em amostras múltiplas. É um processo importante para compreender como as variáveis se interrelacionam e para identificar padrões subjacentes em grandes conjuntos de dados (McEnery, 2001: 89).

Num segundo momento vamos fazer uma análise das sequências dos TD para identificar padrões ou sequências comuns ao longo do *corpus*. O Gráfico 2 mostra as sequências dos TD para os textos 702 e 703 (Anexo 9), usados como exemplo, e que foram selecionados aleatoriamente. Estes textos aparecem sequencialmente no *corpus* e têm, respetivamente, nove e oito sequências de texto com TD diferentes,

[37] O método de ligação de Ward (*Ward's Linkage*) é uma técnica usada para minimizar a variância total dentro de cada cluster (Tan, 1999: 516-517).

[38] O Gráfico 16 com a distribuição dos TD por género foi elaborado com recurso ao *RapidMiner*.

representados pelos pontos no gráfico. O eixo X do gráfico representa as sequências TD (por exemplo, o ponto 0 corresponde ao DI, o ponto 1 ao DT, e assim sucessivamente) ao longo do *corpus* e o eixo Y o tipo discursivo em cada sequência.

O texto 702 começa no o primeiro ponto (ponto 0 do gráfico) em DI, seguido de um DT (ponto 1), depois novamente DI (ponto 2), seguido de um RI, e assim sucessivamente até à sequência número 9 (ponto 8 do gráfico), que é seguido de uma sequência DI, e que corresponde à primeira sequência do texto 703. Podemos, pois, observar que o primeiro texto abre e fecha com uma sequência do tipo Discurso Interativo, enquanto o segundo texto abre com uma sequência Discurso Interativo e fecha com uma sequência de Discurso Teórico.

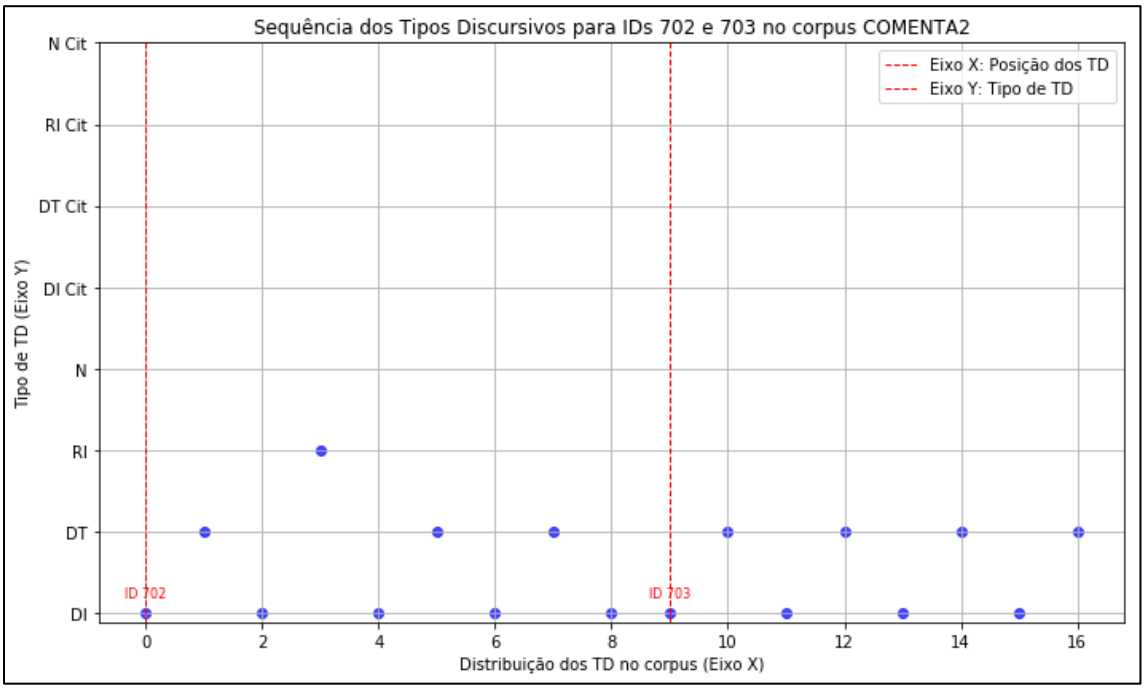


Gráfico 2: Sequência dos TD para os textos 702 e 703.

Finalmente, vamos agrupar os tipos discursivos em subsequências[39], que pretendem mostrar se existem conjuntos de TD que coocorrem com mais frequência no comentário. Esta etapa será feita com recurso a um *script* de *Python* que,

[39] Utilizamos, o termo subsequências para designar combinações ordenadas de elementos extraídas de uma sequência maior. No contexto dos dados que estamos a analisar, uma subsequência é uma combinação de tipos discursivos que aparecem consecutivamente numa sequência associada a um texto/*corpus* específico.

primeiramente, irá somar a frequência das subsequências sem distinguir os textos e, seguidamente, irá ordenar os resultados por frequência. O resultado deste *script* será um gráfico de barras onde serão analisadas as vinte subsequências mais comuns no COMENTA2. As subsequências apresentadas terão um comprimento mínimo de 2 e um máximo de 4, o que permitirá observar a combinação dos quatro tipos discursivos, se necessário. Vejamos um exemplo com as subsequências apresentadas no Gráfico 1:

DI	DT	DI	RI	DI	DT	DI	DT	DI	DI	DT	DI	DT	DI	DT	DI	DT

As subsequências de comprimento 2 seriam:

- "DI DT"
- "DT DI"
- "DI RI"

As subsequências de comprimento 3 seriam:

- "DI DT DI"
- "DT DI RI"

Um segundo gráfico mostrará a distribuição das frequências das subsequências agrupadas por comprimento (2, 3 ou 4 elementos), e mostrará qual o comprimento das subsequências mais comuns. O resultado da análise destas sequências e subsequências pretende mostrar a robustez do padrão de TD enquanto variável para a classificação da variável alvo (no nosso caso, o género textual) no modelo de classificação. Se, como pensamos, for possível caracterizar os TD que ocorrem no *corpus*, seja pela frequência e/ou pelo padrão que revelam ao longo dos textos, parece-nos que pode ser considerada uma característica intrínseca o género textual do comentário.

Na terceira fase deste trabalho, foi realizada uma análise exploratória dos dados, utilizando um operador para ler a folha de cálculo. Esta operação básica permite-nos quantificar os dados em análise e fazer algumas observações como contabilizar e relacionar os atributos de cada classe (veja-se, por exemplo, o Gráfico 3 onde podemos

observar a relação entre os tipos discursivos identificados e o tema do texto a que pertencem):

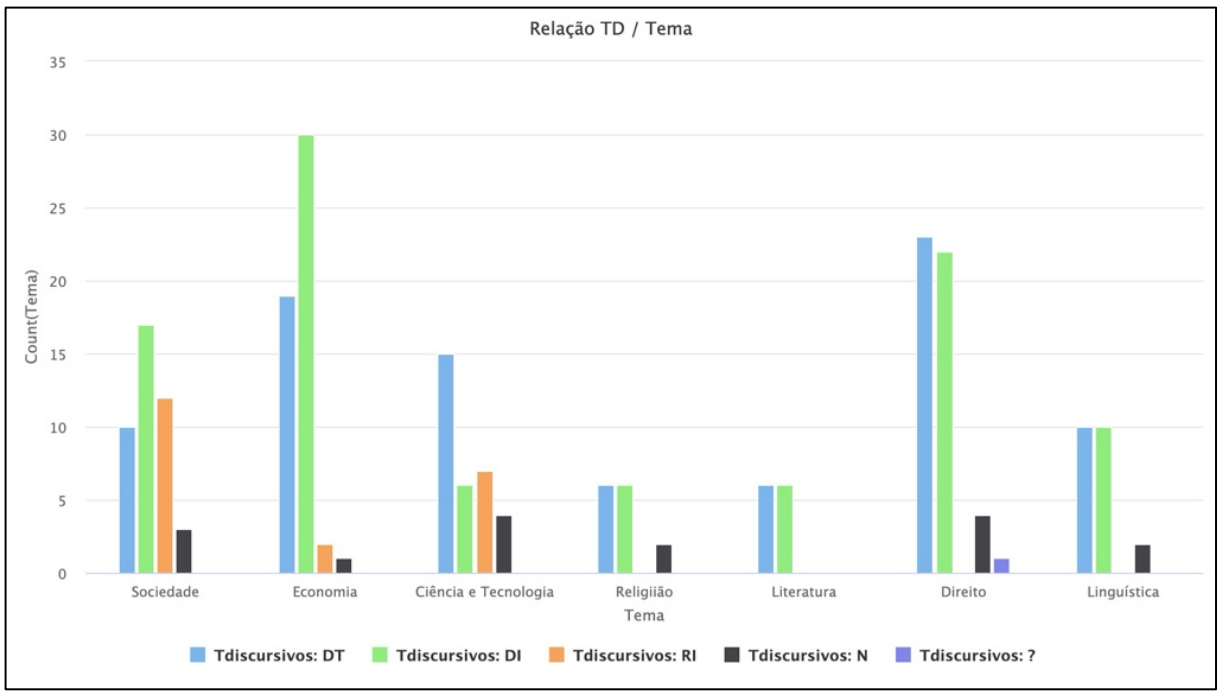


Gráfico 3: Relação entre Tipos Discursivos e Tema

No segundo momento, partimos dos dados obtidos e executámos um *AutoModel*. O *AutoModel* do *RapidMiner* é uma ferramenta integrada que simplifica a criação de modelos de *Machine Learning* e *Data Mining*, que utiliza abordagens automatizadas para pré-processamento, seleção de algoritmos e avaliação dos resultados. Depois de o conjunto de dados ser carregado, o *AutoModel* executa automaticamente uma análise dos dados, identificando tipos de atributos (nominais, numéricos, etc.) das classes, bem como os valores ausentes.

Os dados são analisados e validados pelo *AutoModel*, e o utilizador seleciona a tarefa que pretende executar. O *RapidMiner* permite três tarefas:

- Classificação: Prever uma categoria ou classe. O tipo de previsão pode ser categórico ou numérico contínuo.
- Clusterização: Agrupar dados com base em similaridade.
- *Outliers*: Identificar registos que diferem significativamente da maioria dos dados.

As tarefas de Classificação, que utilizaremos para o nosso modelo, baseiam-se em árvores de decisão baseado no algoritmo C5.0 (aprendizagem supervisionada[40]). Os modelos que recorrem às árvores de decisão aprendem a partir de um conjunto de dados pré-classificados e constroem um modelo, com os padrões encontrados, que utilizam para classificar novos dados (Tan et al., 2018: 491). No nosso caso, faremos uma tarefa de classificação, para executar uma previsão do tipo categórico para o género textual ("Comentário" ou "Notícia"), que tem por objetivo atribuir uma etiqueta (*label*) ao grupo de treino. Depois de os dados serem carregados, o *AutoModel* divide-os em conjuntos de treino e teste (validação cruzada) e treina os modelos, enquanto calcula as métricas de desempenho.

A avaliação do modelo é uma das fases mais importante porque não se trata apenas de decidir qual o modelo que tem o melhor resultado, mas sim aquele que tem melhor resultado para o problema em análise. Para a avaliação do desempenho dos modelos, vamos analisar dados de várias técnicas, das quais destacamos duas:

a) *Confusion matrix*: Baseia-se na contagem dos registos de teste correta e incorretamente previstos pelo modelo (Tan et al., 2018: 149). Estas contagens são apresentadas numa tabela (Tabela 9), e é a partir da *confusion matrix* que obtemos uma série de medidas de avaliação de desempenho do modelo como a *recall*, *precision* e *classification error*[41]. Na Tabela 9 *p* corresponde uma classificação positiva, enquanto *n* corresponde a uma classificação negativa*:* é a interceção destas categorias que constroem a *confusion matrix*. Na nossa análise, vamos

[40] Na aprendizagem supervisionada, existe um supervisor ou mentor externo, que representa o conhecimento do ambiente por meio de conjuntos de dados no formato (entrada, saída). O algoritmo de *machine learning* é então treinado a partir dos conjuntos de exemplos rotulados visando aprender um modelo que permita resolver o problema apresentado. No nosso caso, dado um conjunto *input* (atributos de cada um dos textos), o algoritmo deve encontrar uma regra que nos permita prever o *output* associado a cada novo *input* (o género em que se inscreve).

[41] A *precision* determina a fração de registos que realmente se revelam positivas no grupo que o classificador declarou como uma classe positiva. Quanto maior a precisão, menor o número de erros falsos positivos cometidos pelo classificador. O *recall* mede a fração de exemplos positivos corretamente previstos pelo classificador. Classificadores com grande *recall* têm muito poucos exemplos positivos classificados incorretamente como a classe negativa. Na verdade, o valor do *recall* é equivalente à verdadeira taxa positiva. (Tan et al., 2018)

classificar os textos com base no género. Depois de treinar o modelo com um *corpus* de treino, o modelo aprende a distinguir os dois atributos. Quando o modelo faz previsões, pode cometer quatro erros ou acertos:

- *True Positive*: O modelo prevê "comentário" e, de facto, o texto é um comentário.
- *True Negative*: O modelo prevê "notícia" e o texto é uma notícia.
- *False Positive* : O modelo prevê "comentário", mas o texto é, na verdade, uma notícia. Este erro é conhecido como "falso positivo".
- *False Negative* (FN): O modelo prevê "notícia", mas o texto é um comentário. Este erro é chamado de "falso negativo".

A *confusion matrix* organiza esses valores numa tabela, permitindo uma análise detalhada do desempenho do modelo.

	p	**n**
P	*True Positive*	*True Negative*
n	*False Positive*	*False Negative*

Tabela 7: Confusion Matrix

b) *Cross-Validation*: é uma metodologia de estimativa de erro, em que cada registo é usado o mesmo número de vezes para treino e para testes. Depois de os dados serem divididos nos respetivos conjuntos de tamanho igual, um dos conjuntos é usado para treino e o outro para testes. Em seguida, o papel dos conjuntos é trocado para que o conjunto de treino anterior ocupe a função de teste e vice-versa. O erro total é obtido somando os erros de ambas as execuções (Tan et al., 2018: 187). No caso do *RapidMiner*, a *performance* é calculada com 40% dos dados, que não são utilizados em nenhuma das otimizações realizadas no modelo. Este conjunto é, depois, usado como *input* para a validação de múltiplos conjuntos de reserva (*hold-out*), onde calculamos a *performance* para 7 subconjuntos disjuntos. O maior e melhor desempenho é removido e a média das 5 *performances* restantes é apresentada. Não se trata, portanto, de uma *cross-*

validation total, mas procura equilibrar o tempo de execução e a qualidade de validação do modelo[42].

As técnicas de avaliação aqui apresentadas são adequadas para problemas com duas classes (positiva e negativa) e, no caso da *cross-validation*, têm a vantagem de utilizar o máximo de dados possível para o treino do modelo. No entanto, não existe um consenso sobre qual o melhor classificador, sendo necessário, para isso, testar vários classificadores e comparar as métricas de avaliação (Feldman & Sanger, 2007: 80). A avaliação de um processo de *data mining* envolve não só medir o desempenho de um modelo, mas também considerar a qualidade e a quantidade dos dados disponíveis, os tipos de erros, os custos envolvidos e, em última análise, as implicações filosóficas[43] da interpretação dos resultados. Para a quantidade de dados disponíveis (pequeno volume de dados) e qualidade dos dados (sem atributos em falta), a *cross-validation* é a técnica de avaliação mais adequada (Tan et al., 2018: 187; Witten & Frank, 2005: 149).

O *AutoModel* do *RapidMiner* aplica diversos algoritmos, próprios para a tarefa em questão. Não estando no âmbito deste trabalho explicar como funciona cada um os algoritmos[44], vamos sintetizar as suas funções:

- *Naïve Bayes*: baseia-se no teorema de Bayes, assumindo a independência entre os atributos.
- *Generalized Linear Model* (GLM): expande a regressão linear para lidar com variáveis de saída (resposta) que não seguem a distribuição normal, permitindo trabalhar dados categóricos, de contagem ou contínuos.
- *Logistic Regression*: modelo estatístico que estima a probabilidade de um exemplo pertencer a uma classe (binária ou multiclasse).

[42] https://docs.rapidminer.com/2024.1/studio/operators/validation/cross_validation.html

[43] Referimo-nos a questões como a natureza da Ciência e o conceito de avaliação (Witten & Frank, 2005: 144).

[44] Os algoritmos usados em *Text Mining* e *Data Mining* são objeto de revisões e análises periódicas. Para uma descrição detalhada do funcionamento e objetivo de cada um dos algoritmos, consultar Witten & Frank, (2005).

- *Fast Large Margin* (FLM): Similar ao SVM, está otimizado para conjuntos de dados grandes, priorizando a eficiência.
- *Deep Learning* (DL): rede neural profunda com várias camadas, capaz de capturar padrões complexos nos dados.
- *Decision Tree*: Divide os dados em grupos com base em condições simples, formando uma estrutura hierárquica semelhante a uma árvore.
- *Random Forest*: Combina várias árvores de decisão (floresta) para aumentar a precisão e reduzir o *overfitting*.
- *Gradient Boosted Trees*: Combina vários modelos fracos (como árvores de decisão) sequencialmente, otimizando o desempenho em cada etapa.
- *Support Vector Machine* (SVM): separa dados em categorias diferentes encontrando a "linha" ou "fronteira" que melhor divide os atributos.

O *RapidMiner* executa automaticamente uma avaliação aos vários algoritmos, distinguindo aqueles que têm melhores resultados. O algoritmo que se destacou e obteve os melhores resultados (*Best Performance*, *Best Gain* e *Fastest Scoring Time*) foi o GLM. Vamos, por isso, explicar brevemente como funciona este algoritmo.

O GLM (ou outro modelo de Regressão Linear) é um modelo adequado para a análise de atributos numéricos ou nominais. O fundamento deste modelo (Witten & Frank, 2005: 119-120) é expressar a variável como uma combinação linear dos atributos, com pesos predeterminados, de acordo com a fórmula seguinte:

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots$$

em que x é a variável de saída (numérica), a é o atributo e w é o peso. O peso de cada atributo é calculado com o conjunto de treino, sendo para isso necessário converter os atributos nominais em atributos numéricos (*dummy variable*) (James, Witten, Hastie, & Tibshirani, 2022: 83).

Por exemplo, para a variável Tipo Discursivo (TD), com quatro atributos, o modelo cria três atributos numéricos binários:

- TD_1 : 1 se for TD1, 0 caso contrário.
- TD_2 : 1 se for TD2, 0 caso contrário.

- TD_3: 1 se for TD3, 0 caso contrário.
- TD_4 é a categoria de referência (o que significa que todas os *dummies variables* serão 0 se a observação pertencer a TD4).

Cada uma das variáveis será transformada do mesmo modo, sendo que um dos atributos será sempre a categoria de referência:

- Género: codificado como G_1, G_2 (se houver três categorias, uma servirá de referência).
- Atividade: codificado como $A_1, A_2, \ldots, A_n$.
- Tema: codificado como $T_1, T_2, \ldots, T_n$.
- Tipos Discursivos (TD): TD_1, TD_2, TD_3.

Depois de o modelo aprender os pesos (coeficientes) de cada atributo durante o treino, ele insere os valores *dummies* (0 ou 1) na fórmula.

Exemplo: Suponhamos que o modelo estabeleceu os seguintes coeficientes durante a fase de treino:

Coeficiente	Valor
w_0	50
w_1	10
w_2	5
w_3	12
w_4	8
w_5	15
w_6	7
w_7	6
w_8	4
w_9	2

Como exemplo vamos usar a seguinte linha (hipotética):

G1	G2	A1	A2	T1	T2	TD1	TD2	TD3
1	0	1	0	1	0	1	0	0

Substituímos a fórmula apresentada pelos valores:

$$x = w_0 + w_1 * G1 + w_3 * A1 + w_5 * T1 + w_7 * TD1$$

$$x = 50 + (10 * 1) + (12 * 1) + (15 * 1) + (6 * 1)$$

$$x = 50 + 10 + 12 + 15 + 6 = 85$$

Sendo que, para este exemplo, a classe de saída seria 85. No contexto de um modelo de classificação, a classe de saída representa a categoria atribuída à observação analisada. Esta é obtida através de uma combinação ponderada das variáveis de entrada do modelo, onde os pesos associados refletem a importância relativa de cada característica na previsão final. O modelo prevê a classe com maior pontuação, indicando a categoria mais provável para o objeto analisado, com base na aprendizagem durante o treino. Se, hipoteticamente, o resultado das pontuações para cada classe (Notícia ou Comentário) for Notícia: 85 pontos e Comentário: 65 pontos, o modelo prevê que o texto pertence à classe “Notícia”, pois é a categoria com a pontuação mais alta.

A vantagem deste modelo é que basta fornecermos mais exemplos do que atributos para podermos executá-lo com algum grau de fiabilidade. Mas tem como desvantagem, como o próprio nome indica, o facto de ser linear, o que significa que o modelo calcula um valor chamado previsor linear, que é uma soma ponderada (linear) dos atributos. Para executar a soma ponderada, o modelo parte do princípio de que os atributos têm sempre uma relação linear. Se a dependência for não-linear, o modelo procurará a relação mais direta (Witten & Frank, 2005: 120).

5. Análise dos dados

5.1. Distribuição das categorias sintáticas

Na primeira fase analítica do nosso trabalho, vamos fazer uma caracterização geral das unidades linguísticas, comparando os *subcorpora* CETEM, COMENTA2 e COMJUR. Começamos pela distribuição das categorias sintáticas, no Gráfico 4.

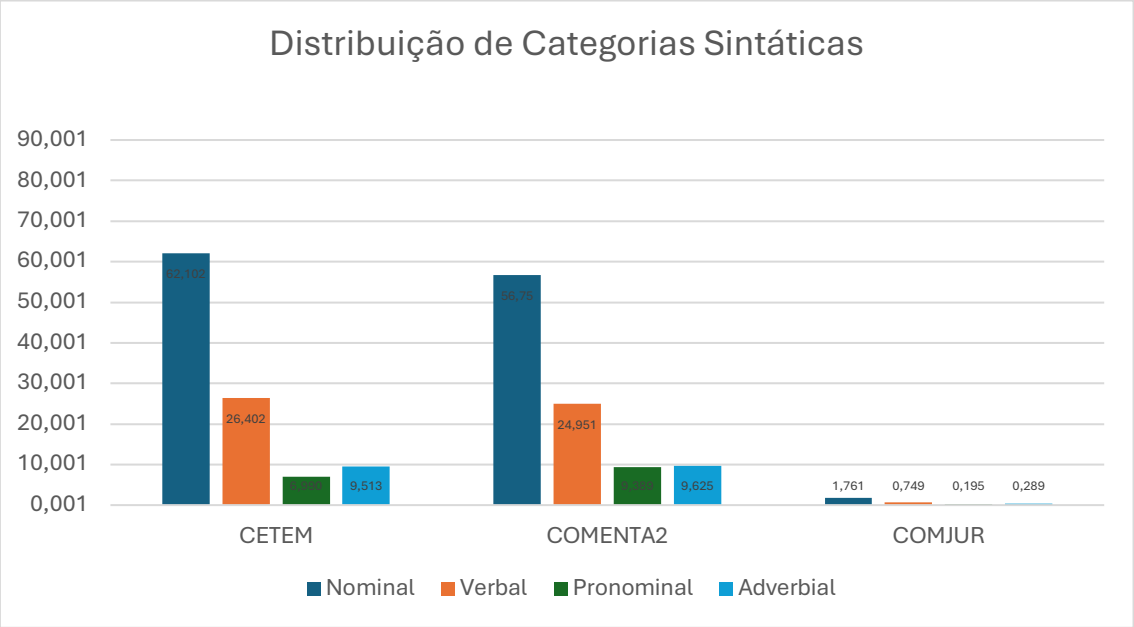


Gráfico 4: Distribuição das categorias sintáticas pelos subcorpora.

De acordo com o gráfico obtido, existe uma semelhança entre os diversos *subcorpora* (CETEM e COMENTA2) com valores muito próximos. O COMJUR surge aqui destacado pela diferença de tamanho (os dados apresentados são absolutos). Destaca-se, como única diferença relevante, o número de pronomes usado no COMENTA2, ligeiramente superior, e que se destaca comparativamente ao CETEM, com uma diferença percentual de 34,39%. O uso de pronomes é uma das poucas diferenças ao nível das unidades linguísticas que conseguimos identificar, como veremos adiante.

Para analisarmos a probabilidade de significância entre os diferentes conjuntos de dados, recorremos a um teste ANOVA de uma via, e os resultados dos testes foram os seguintes:

F-valor	2.186
P-valor	0.168

Tabela 8: Significância estatística (Categorias gramaticais).

O valor-p de 0.168 indica que não há evidências estatisticamente significativas para rejeitar a hipótese nula ao nível de significância comum (geralmente 0.05). Significa, portanto, que não podemos concluir que há uma diferença estatisticamente significativa na distribuição das categorias sintáticas observadas entre os três conjuntos de dados (CETEM, COMENTA2, COMJUR), e que as diferenças observadas entre os conjuntos não são estatisticamente significativas.

5.2. Pronomes

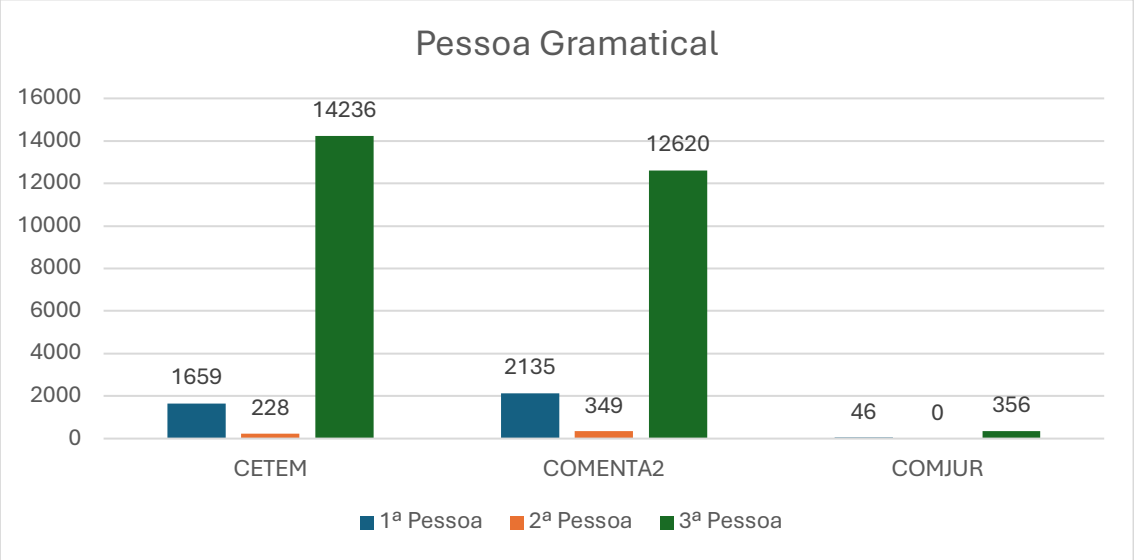


Gráfico 5: Pessoa gramatical presente no corpus.

No Gráfico 5, podemos observar um aspeto morfossintático dos verbos que é a flexão verbal de pessoa. Há uma clara preferência pelo uso da terceira pessoa (tanto no singular como no plural), seguido pelo uso da primeira pessoa e finalmente da segunda pessoa. Existe uma diferença entre os vários *subcorpora*, com o COMENTA2 a revelar uma diferença relativamente ao CETEM: o COMENTA2 mostra uma variação percentual, relativamente ao uso da 1ª pessoa, de 28,69%; e de 53,07% em relação à 2ª pessoa.

F-valor	0.748472
P-valor	0.512626

Tabela 3: Significância estatística (pessoa sintática).

De acordo com o teste ANOVA realizado, os resultados mostram que, com base no valor-p encontrado (0.512626), não há evidências estatisticamente significativas para afirmar que há diferenças entre as médias dos grupos. Isto sugere que, com esta amostra e métodos usados, não podemos concluir que os grupos diferem em termos das variáveis medidas.

No Gráfico 6 podemos observar as ocorrências dos pronomes de acordo com a função sintática que ocupam. Para esta recolha, foram identificados os pronomes, de 1ª e 2ª pessoa, com função de sujeito, de complemento direto, de complemento oblíquo e de agente da passiva.

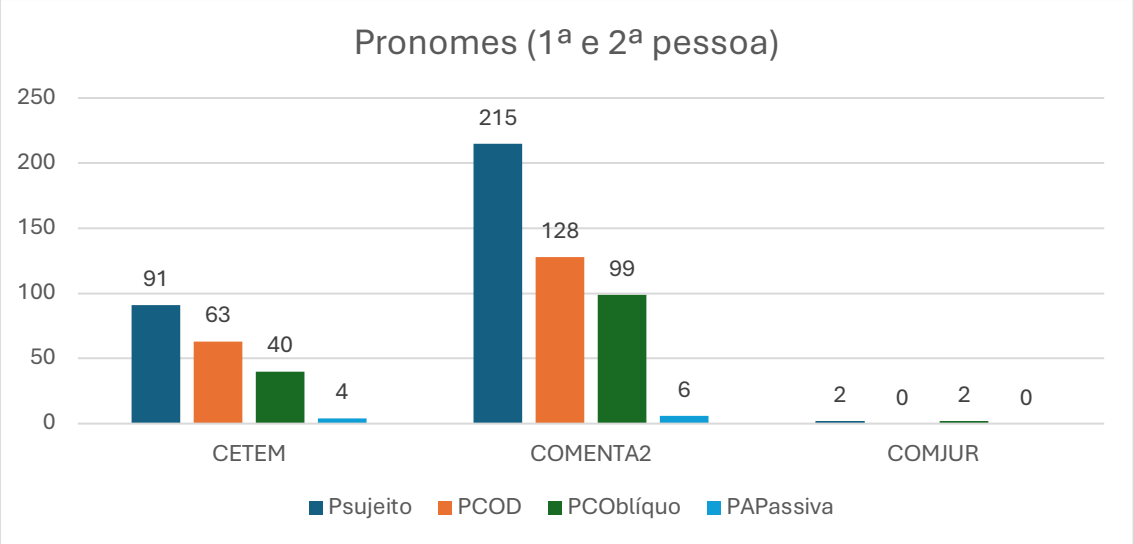


Gráfico 6: Distribuição dos pronomes por função.

Os dados do Gráfico 6, mostram que é ao nível dos pronomes que existem mais diferenças. Há uma diferença significativa entre o uso de pronomes no *subcorpus* COMENTA2 em relação ao *subcorpus* CETEM. Assim, os pronomes de sujeito têm uma variação percentual de 136,26% e os pronomes de Complemento Direto de 103,17%. Os pronomes de Complemento Oblíquo têm uma variação percentual de 147,5% e os pronomes com função de Agente da Passiva têm uma variação de 50%. A variação percentual é resumida na Tabela 3:

	CETEM	COMENTA2	Variação %
Psujeito	91	215	136,26%
PCOD	63	128	103,17%
PCOblíquo	40	99	147,5%
PAPassiva	4	6	50%

Tabela 9: Pronomes com possível interpretação deítica.

Estes dados estão em concordância com o resultado do teste ANOVA:

F-valor	2.508085
P-valor	0.0506

Tabela 10: Significância estatística (Pronomes)

Com um valor-p de aproximadamente 0.0506, há uma tendência em direção à rejeição da hipótese nula, sugerindo que pode haver diferenças significativas entre as médias dos grupos CETEM, COMENTA2 e COMJUR. No entanto, é importante ressalvar que 0.0506 é ligeiramente superior que 0.05, o que significa que os resultados não são conclusivamente significativos a um nível de 5% de significância. Com um valor-p de aproximadamente 0.0506, podemos dizer que há uma indicação de diferença entre as médias dos grupos, mas não podemos afirmar com certeza estatística completa, dado o nível de significância adotado (0.05). Se o contexto permitisse um nível de significância ligeiramente maior, como 0.10, poderia ser considerado um resultado significativo.

Esta diferença é particularmente importante para a identificação e análise dos tipos discursivos. De facto, para Bronckart (2008) os pronomes pessoais, especialmente aqueles com valor deítico, estabelecem uma relação com os tipos discursivos, na medida em que influenciam a organização do discurso, estabelecendo uma relação entre as instâncias de agentividade do texto e as instâncias externas da ação linguística (agente produtor, interlocutor, espaço e tempo de produção), que pode ser de autonomia ou de implicação. A presença de um elevado número de pronomes (de primeira e segunda pessoa) no COMENTA2 mostra uma convergência entre as instâncias agentivas no texto

e as instâncias externas da situação de comunicação (como o produtor e os seus potenciais recetores).

Vejamos, agora, alguns exemplos retirados dos *subcorpora*:

(i) *Quero **eu** dizer que se trata duma duplicação irónica do próprio acto de necrófila arqueologia em que a escritora se vai envolver, ao apropriar-se de objectos do passado duma cultura estranha.* (COMENTA2)

(ii) *Na primeira quadra, o enunciador dirige-se a um **tu**, que é (parece) ridicularizado, satirizado; na segunda, um **eu** fala de si com tristeza, lamentando a sua solidão*. (COMENTA2)

(iii) " *Fomos passear à Serra, **eu** , o meu marido , a minha filha , o Marcelo e o meu enteado , e um dia que podia ser perfeito transformou-se de repente num pesadelo*." (CETEM).

Os exemplos mostram que os valores dos pronomes pessoais são diferentes e a sua interpretação depende do contexto em que ocorrem. No exemplo (i) encontramos um pronome pessoal com função de sujeito que implica o agente com os parâmetros físicos da ação, embora os tempos verbais não tenham um valor deítico, configurando um Discurso Interativo. No segundo exemplo, "eu" e "tu" são pronomes masculinos (como podemos confirmar pela presença dos artigos indefinidos) e que servem aqui para mostrar como os etiquetadores automáticas podem alterar as métricas. Neste caso, a atividade (académica) e o tema do texto (literatura) são importantes para a leitura dados. O último exemplo ocorre num contexto de transcrição do discurso direto, próprio do género notícia, e configura um narrar implicado.

5.3. Tempos e Modos verbais

Os modos verbais utilizados nos *subcorpora* revelam uma preferência pelo modo Indicativo, seguido de formas nominais dos verbos como o Particípio Passado e o Infinitivo, podendo estes últimos estarem associados ao uso de tempos compostos e de construções perifrásticas. Por outro lado, o *subcorpus* COMENTA2 mostra uma preferência pelo uso do conjuntivo, quando comparado com o CETEM, numa variação percentual de 10,78% e de 42,56% para o uso do gerúndio.

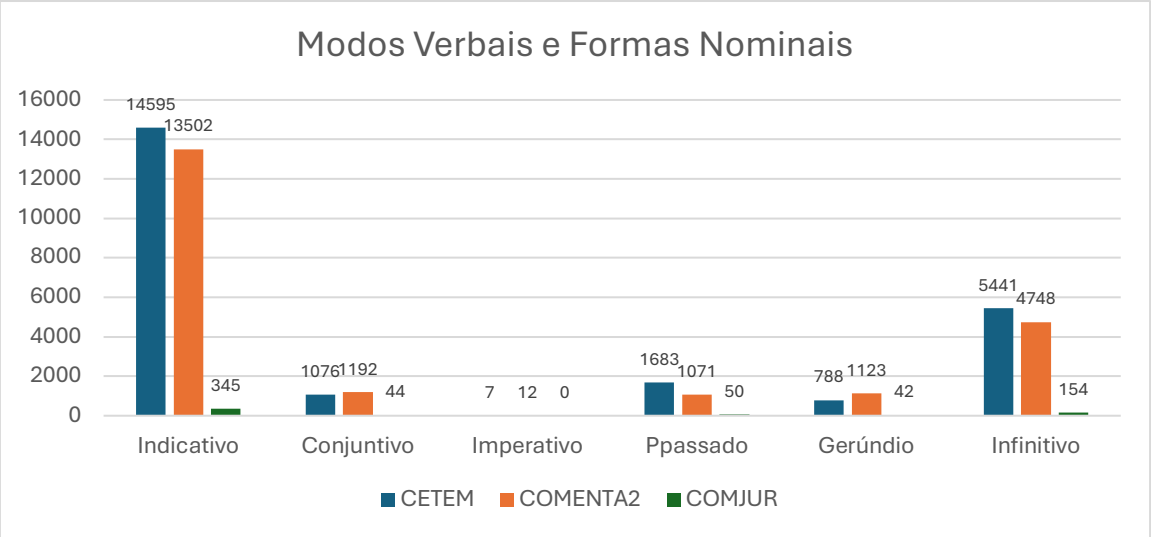


Gráfico 7: Modos verbais e formas nominais.

O uso do modo Conjuntivo está tradicionalmente (Mateus, Brito, Duarte, & Faria, 2006) associado a construções de frases coordenadas e subordinadas, em que o seu uso (obrigatório ou opcional) apresenta valores diferentes, como incerteza, eventualidade ou dúvida. O modo Indicativo está a associado a valores de certeza, em construções de frases simples, coordenadas e como tempo verbal preferencial de orações subordinantes. Se não podemos analisar todos os valores que os modos Indicativos e Conjuntivos adquirem nos textos dos nossos *corpora*, podemos, pelas métricas obtidas, associar sintaticamente o Conjuntivo a estruturas frásicas mais complexas. Dentro do quadro do ISD, Bronckart (1997, 2008) refere os modos indicativo e conjuntivo na relação que estabelecem com a modalização de expressões temporais e a organização dos discursos: as modalizações são realizadas por diferentes unidades ou estruturas, incluindo os tempos verbais, que podem ser categorizados em modos, como o indicativo e o condicional. O modo indicativo é geralmente associado a afirmações de factos e realidades, enquanto o modo condicional (que pode ser considerado uma forma de modo conjuntivo) expressa possibilidades, hipóteses ou condições. Bronckart destaca que as marcas de modalização não se organizam em séries isotópicas, mas são frequentemente sobrepostas ou “inseridas” dentro de estruturas que têm outras funções, como os tempos verbais, verbos auxiliares e advérbios, e que a análise das modalizações deve considerar a natureza das operações subjacentes, que envolvem julgamentos "meta" aplicados à ordem temática, transcendente ao tipo de discurso. Isto

significa que a escolha entre modos verbais pode influenciar a forma como os discursos são estruturados e compreendidos.

Relativamente à significância estatística, a ANOVA mostra que, com um valor-p de aproximadamente 0.1821, não há evidências estatisticamente significativas para afirmar que há diferenças nas médias dos grupos.

F-valor	1.4207
P-valor	0.2722

Tabela 11: Significância estatística (Modos Verbais)

O valor de *F* de 1.911829, embora sugerindo alguma variabilidade entre os grupos, não é suficientemente grande para ser considerado significativo ao nível de 0.05. Estes resultados sugerem que, com base nos dados fornecidos e ao nível de significância de 0.05, não podemos concluir que os grupos diferem significativamente em termos das variáveis medidas, e que as diferenças observadas nas médias dos grupos podem ser atribuídas a escolhas do produtor do texto.

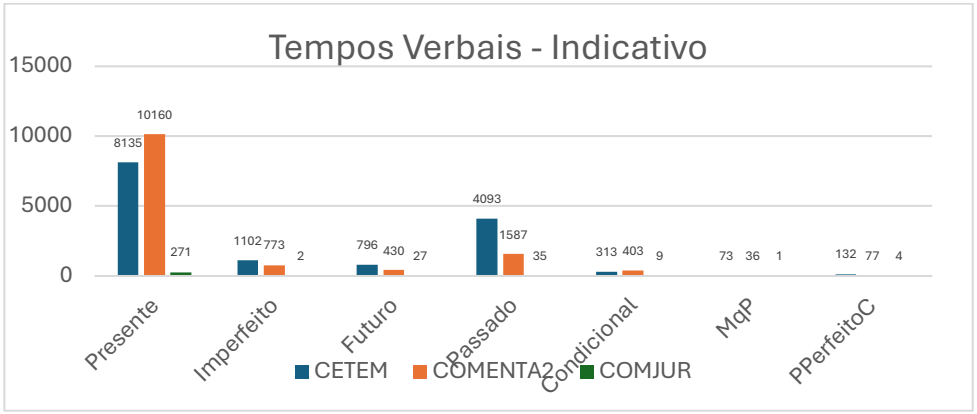


Figura 18: Tempos verbais do Modo Indicativo.

O resultado da ANOVA sugere que os grupos não diferem significativamente em termos das variáveis medidas.

F-valor	1.198323
P-valor	0.324656

Tabela 12: Significância estatística (Tempos Verbais do Indicativo)

O valor-p de 0.324656 é significativamente maior que o nível de significância comum de 0.05, enquanto o valor de F de 1.198323, tal como nos testes anteriores, sugere alguma variabilidade entre os grupos, mas não é suficientemente grande para ser considerado significativo ao nível de 0.05. Nos *subcorpora* destaca-se o uso de tempos no Indicativo, com preferência pelo Presente do Indicativo e pelo Pretérito Perfeito Simples. Observamos, também, que o COMENTA2 tem uma preferência pelo Presente, enquanto o CETEM tem mais ocorrências de verbos no Pretérito Perfeito Simples.

De acordo com Bronckart (1997, 2008), a organização temporal é um dos aspetos mais relevantes para a análise dos TD, uma vez que o autor critica as abordagens gramaticais tradicionais que consideram que os valores de temporalidade são expressos apenas pelos determinantes dos verbos (tempos verbais) e em interação com alguns advérbios. Bronckart propõe que a análise temporal das relações entre os tempos dos verbos e os momentos de produção e do processo verbal seja mais detalhada do que as abordagens tradicionais, propondo uma análise com três parâmetros:

> (i) O momento de produção que integra a duração psicologicamente construída em torno do ato de produção linguística e não um simples ponto físico ou objetivo de produção.
>
> (ii) O momento do processo verbal que está relacionado com o evento ou ação que o verbo descreve.
>
> (iii) O momento psicológico de referência que se encontra relacionado com o contexto temporal no qual o discurso é situado, como por exemplo, um tempo expresso por adjetivos de tempo como “hoje”, “amanhã” ou “ontem”.

Deste modo, a relação entre o momento de produção, o processo verbal e o momento de referência é fundamental para entender o valor dos tempos verbais no ISD. Por exemplo:

- O presente marca uma relação de simultaneidade entre o processo e o momento de referência.
- O futuro indica uma relação de posterioridade, onde o processo acontece depois do momento de produção.

- O passado composto[45] expressa uma relação de anterioridade, em que o processo ocorre antes do momento de produção.

Os tempos verbais assumem, por isso, um lugar de destaque na análise e identificação dos TD, uma vez que, como referimos anteriormente, a temporalidade é um dos eixos em torno do qual se organizam os TD. Assim, no discurso da ordem do Narrar, a relação entre os eventos relatados e a produção do discurso é mais distanciada, com uma origem temporal frequentemente explícita, e com o uso de diferentes tempos verbais para organizar a cronologia dos eventos. Já no discurso da ordem do Expor, não há uma disjunção clara entre as coordenadas temporais do discurso e da produção; os tempos verbais podem ser mais complexos, articulando relações de anterioridade, simultaneidade e posterioridade de acordo com a duração psicologicamente construída da produção.

Nos *corpora* usados na nossa análise, os tempos verbais utilizados podem estar associados à natureza dos textos, em que a organização temporal de relatos de eventos passados é própria da notícia. Os exemplos de presente do indicativo parecem estar associados a um presente com valores aspetuais diferentes, como podemos ver nos exemplos seguintes:

(i) *dois irmãos norte-americanos de origem francesa, **chegam** depois de amanhã ao FC Porto* CETEM

(ii) *Sidney, de 22 anos e 1,94 metros , **joga** a base-extremo e terminou este ano a Universidade de Princeton* CETEM

(iii) - ***Sou** contrabaixista. **Estou** a cantar por acidente*. CETEM

No exemplo (i) o presente do indicativo tem um valor de posteridade através da expressão temporal "depois de amanhã", enquanto no exemplo (ii) o verbo "joga" tem um valor de simultaneidade, tal como os verbos *Ser* e *Estar* no exemplo (iii).

No COMENTA2, observa-se, no entanto, o uso do presente do indicativo com valores diferentes:

[45] Partindo do exemplo da língua francesa.

(iv) *Está em causa saber-se que a literatura não se **faz** de boas intenções, mas sim de uma objectividade possível, centrada na historicidade dos textos, à luz da importância simbólica e ideológica de que **são** dotados numa determinada cultura*. COMENTA2

(v) *torna-se-nos claro que nem sempre a questão do gosto **acompanha** a questão da justiça e justeza quanto ao corpus seleccionado*. COMENTA2

No exemplo (iv) encontramos verbos no presente do indicativo, com valor gnómico, associado a um discurso do tipo teórico. E no exemplo (v), a implicação presente no pronome "nos" associa-se a um discurso interativo. Estes breves exemplos mostram-nos que uma análise estatística dos tempos verbais não é suficiente para determinar o tipo discursivo, mas parece mostrar algumas diferenças: no CETEM, o presente é menos utilizado e tem valores temporais específicos associados à organização temporal do texto, enquanto no COMENTA2 surge associado a outros valores.

5.4. Advérbios de tempo e lugar

O Gráfico 8, relativo aos advérbios de tempo e de lugar, com possível interpretação deítica, mostra algumas diferenças entre os *subcorpora*. No CETEM há uma clara predominância dos advérbios de tempo que podem estar associadas ao facto de o CETEM ter sido produzido dentro da atividade jornalística, sendo esta diferença explicável pela natureza do texto jornalístico que necessita de ordenar temporalmente os acontecimentos (Miranda, 2010).

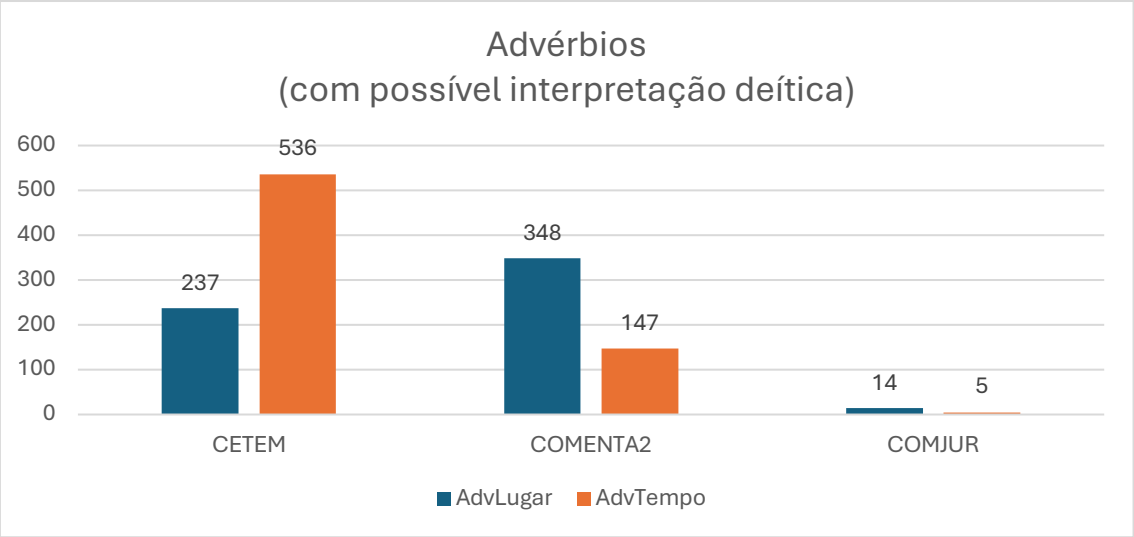


Gráfico 8: Distribuição dos advérbios de lugar e de tempo.

Vejamos alguns exemplos:

(i) *Fernando Henrique Cardoso fala* ***hoje*** *em Brasília* CETEM

(ii) *Foi assim o dia de* ***ontem*** *do Benfica na ressaca da derrota em Vila do Conde*. CETEM

(iii) *ansiedade que tem muito a ver com o "sofrimento do professor" descrito* ***ontem*** *por Ana Cristina Silva*. COMENTA

(iv) *" Vamos* ***agora*** *tomar decisões de fundo - e com suporte legal "* CETEM

(v) *Mas voltemos,* ***agora****, à investigação de cariz cognitivo*. COMENTA

(vi) *Já* ***agora****, refira-se que na editora do disco, a Mute, milita um lote numeroso de " industriais " e outros terroristas da electrónica...* CETEM

(vii) *Um problema que* ***hoje*** *se coloca com acuidade é o de como ensinar o Português Europeu* COMENTA

(viii) *o leitor poderá* ***agora*** *dispor de um instrumento único em O Género Intranquilo*. COMENTA

(ix) *Afinal,* ***ontem*** *como hoje, tudo o tempo corrói na sua inexorável efemeridade* COMENTA

Nos exemplos (i) a (iii) encontramos advérbios de tempo com uma função deítica, localizando temporalmente os factos noticiados, em que o localizador de origem se situa

na data de produção ou publicação do texto. Já os exemplos (iv) e (v) só podem ser interpretados anaforicamente, referindo-se a informação dada anteriormente no texto. Nos exemplos (vi) a (ix), os advérbios integram locuções (vi), expressões idiomáticas (ix) ou veiculam um valor temporal indefinido (vii e viii). Como referimos anteriormente, a temporalidade é um dos aspetos relevantes para a identificação e análise dos TD.

Bronckart (2008) sublinha que os advérbios de tempo desempenham um papel importante na estruturação do discurso e que a análise da temporalidade deve considerar não apenas os tempos verbais, mas também a forma como os advérbios de tempo e lugar se articulam com esses tempos verbais. Comparando, a partir dos exemplos, o valor dos advérbios de tempo entre os *corpora*, no CETEM o uso mais recorrente parece estar associado a um valor deítico, enquanto no COMENTA2 tem um valor indefinido. A ausência de unidades que remetam ao espaço-tempo da produção é uma característica do discurso da ordem do expor implicado, enquanto a presença de advérbios com valor temporal pertence à ordem do narrar. Isso significa que uma análise estatística dos advérbios por si só não é suficiente para identificar os tipos discursivos; é necessário interpretá-los considerando uma dimensão praxeológica. O entendimento do contexto permite examinar e interpretar os dados de um ângulo que a estatística não permite.

As locuções de tempo analisadas no *corpus* do CETEM são mais expressivas em termos quantitativos, uma vez que, das locuções analisadas, apenas seis ocorrências pertencem ao *subcorpus* COMENTA2:

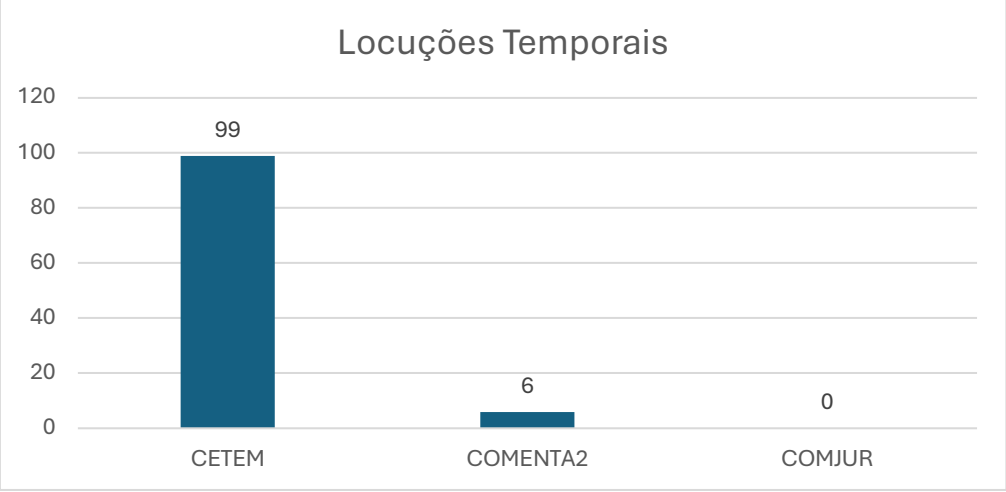


Gráfico 9: Locuções Temporais.

De acordo com a ANOVA feita com os dados relativos ao Gráfico 8, com um valor-p de aproximadamente 0.1716, não há evidências estatisticamente significativas para afirmar que há diferenças nas médias dos grupos.

F-valor	3.358314
P-valor	0.171556

Tabela 13: Significância estatística (Advérbios)

O valor de F de 3.358314, sugere alguma variabilidade entre os grupos, mas não é suficientemente grande para ser considerado significativo.

Os advérbios de lugar não têm, tal como os advérbios de tempo, uma interpretação objetiva. Como vários trabalhos mostram (Correia & Pereira, 2014; Pereira, 2009; Teixeira, 2005), os advérbios de lugar só têm uma interpretação locativa em determinados contextos. A Tabela 16 regista as ocorrências dos advérbios de lugar no *corpus*:

	CETEM	**COMENTA2**	**COMJUR**
aqui	48	159	7
aí	27	35	2
ali	30	26	0
cá	11	10	0
lá	57	33	0
acolá	2	0	0
além	62	85	5

Tabela 14: Ocorrências de advérbios de lugar no corpus.

A tabela mostra-nos que o COMENTA2 tem o maior número de ocorrências de advérbios de lugar, seguido do CETEM e, por último, temos o COMJUR que tem um número muito baixo de ocorrências (vários com o valor 0). Se agruparmos estes valores num gráfico, podemos ver que advérbios são mais comuns em cada um dos *subcorpora*:

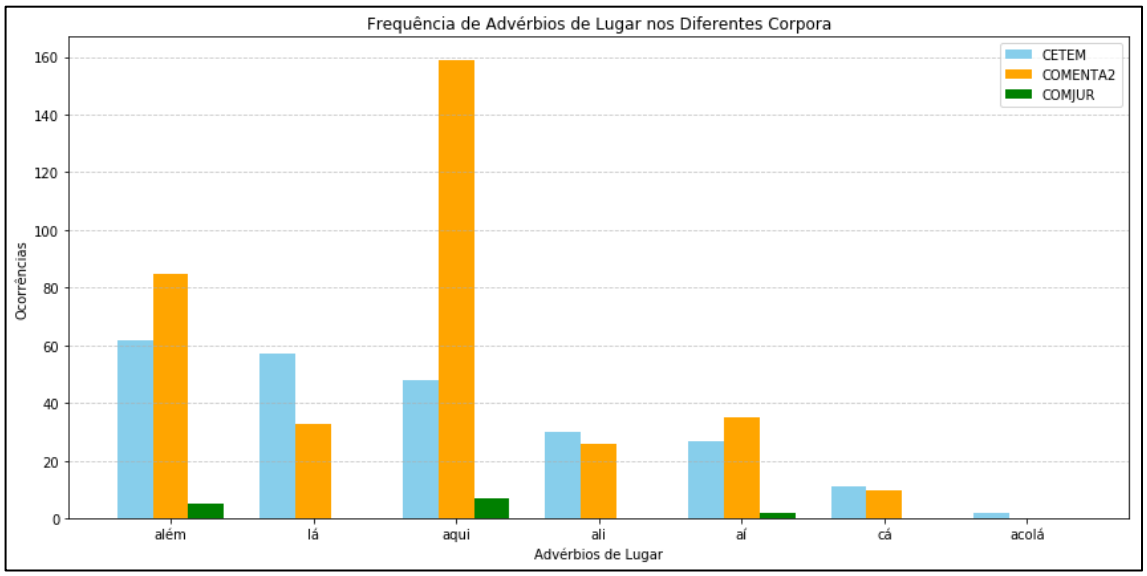


Gráfico 10: Distribuição dos advérbios de lugar pelos diferentes subcorpora.

De acordo com o Gráfico 10, a distribuição advérbios que ocorrem em cada *subcorpus* é:

- **CETEM**: *além, lá, aqui*
- **COMENTA2**: *aqui, além, aí*
- **COMJUR**: *aqui, além, aí*

Estes dados mostram-nos que há uma semelhança no tipo de unidades linguísticas utilizados no COMENTA2 e COMJUR. Embora estes advérbios possam assumir vários valores, as escolhas destas unidades linguísticas parecem ser semelhantes.

(i) *E por **aí** se adivinham as possíveis razões para o facto* COMENTA2

(ii) *dado que, até **aí**, só recebera lições do companheiro*. COMENTA2

(iii) *e **aí** aprofundar uma convivência que a leitura destas páginas lhe tornou grata*. COMENTA2

(iv) *um trabalhador empregado no território de outro Estado-Membro têm o direito de **aí** se instalarem* COMJUR

Os exemplos retirados dos *subcorpora* COMENTA2 e COMJUR, mostram que o advérbio de lugar "aí" assume vários valores: em (i) integra uma locução adverbial de lugar com valor indefinido, em (ii) assume um valor temporal e, em (iii) e (iv) tem um

valor anafórico. Se relacionarmos estes dados com o Gráfico 8, que mostrava que os *subcorpora* COMENTA2 e COMJUR tinham mais ocorrências de advérbios de lugar, podemos dizer que embora haja mais advérbios de lugar, estes não têm uma função deítica. Isto reforça o que dissemos anteriormente quando analisámos os advérbios de tempo: nestes *subcorpora* ocorre uma rutura com o espaço-tempo da produção.

Vejamos mais alguns exemplos com os advérbios *cá* e *lá* retirados do *corpus*.

(i) *Mais tarde, compreenderei que deve ter ido telefonar, para o Director, para o hospital, para o comissário, sei* ***lá*** *para onde mais!* CETEM

(ii) *E também há que compreender o carácter universal que o inglês assume quando se trata de informática ou de assuntos relacionados com a Internet (****cá*** *está ...)* CETEM

(iii) *e* ***cá*** *vou vivendo de gaiola até à remissão dos pecados.* COMENTA 2

(iv) *Reinava então* ***cá*** *no burgo o eng. Santos e Castro de saudosa memória*. CETEM

(v) *As pessoas* ***cá*** *da Amadora são boa gente*. CETEM

(vi) *Não é por importância, mas por verificar - através das reacções e das atitudes em Portugal ou* ***lá*** *fora, no estrangeiro - que transporto comigo uma certa responsabilidade ...* CETEM

Podemos observar nos exemplos que os advérbios de lugar *cá* e *lá* nem sempre têm uma interpretação espacial, como nos exemplos (i), (ii) e (iii), enquanto nos exemplos (iv), (v) e (vi), têm uma interpretação locativa[46]. Significa, portanto, que a análise quantitativa dos advérbios de lugar selecionados não é suficiente para fazer uma representação da *deixis* espacial no *corpus*, mas que existe uma ligeira correlação entre as unidades linguísticas escolhidas e o género a que pertence o texto. Esta correlação será explorada na secção seguinte através de um teste de *clustering*.

[46] Cf. Correia & Pereira (2014).

5.5. Testes de *clustering*

Com base nos resultados dos testes ANOVA que foram realizados anteriormente, não há evidências suficientes para concluir que existam diferenças estatisticamente significativas entre os grupos analisados. Isto é indicado pelos valores-p relativamente altos (todos acima de 0.05), o que revela que a diferença entre as médias dos grupos não pode ser considerada intencional, dificultando a identificação de padrões claros baseados nas categorias analisadas. No entanto, estes resultados não invalidam a realização de técnicas de *clustering*, que possam identificar grupos ou padrões nos dados com base em similaridades gerais, e não necessariamente em diferenças significativas das médias.

5.5.1. *Hierarchical Clustering*

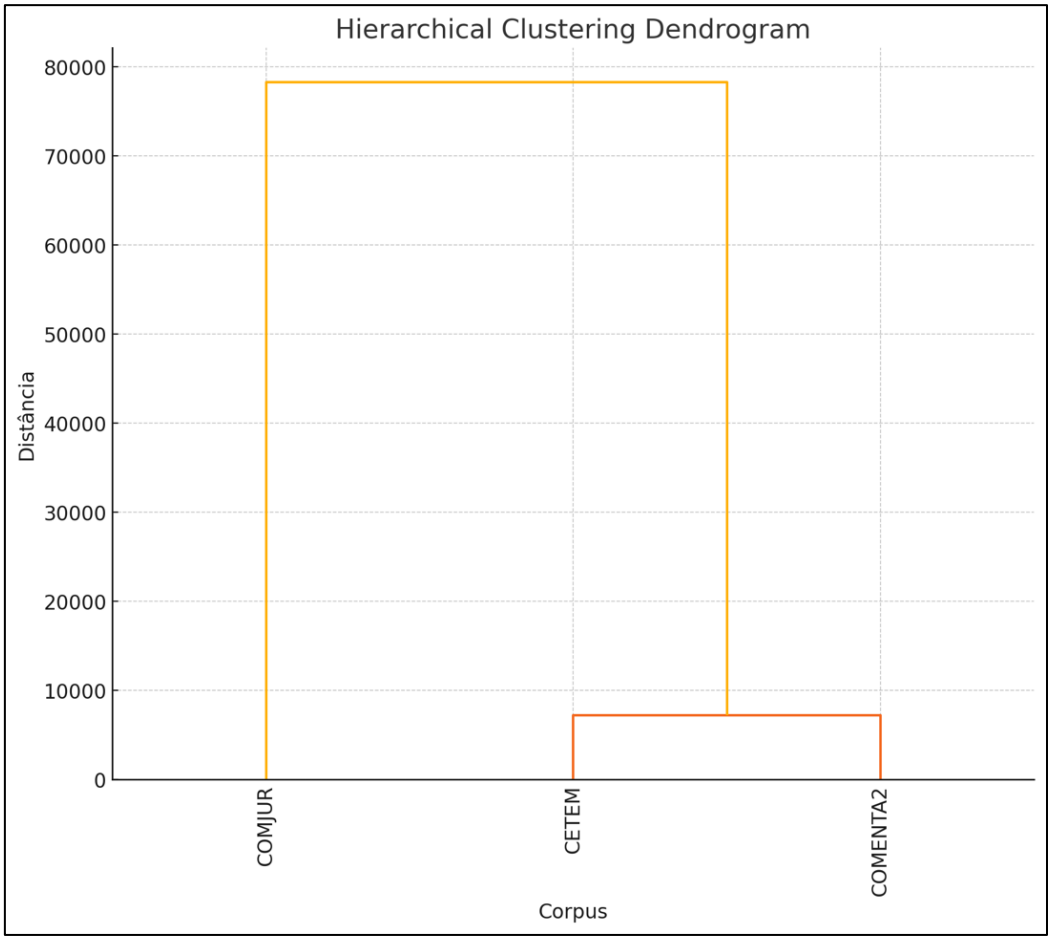


Gráfico 11: Dendrograma dos subcorpora.

De acordo com o dendrograma obtido (Gráfico 11), o *subcorpus* COMJUR é significativamente distinto dos outros dois (CETEM e COMENTA2). O COMJUR forma um

cluster separado, com uma distância muito maior, o que indica que as suas características são diferentes em relação aos outros. Já os *subcorpora* CETEM e COMENTA2 estão mais próximos um do outro, formando um *cluster* a uma distância muito menor, o que sugere que compartilham muitas características semelhantes, em comparação com o COMJUR. A altura na qual as linhas se unem, no Gráfico 11, indica a distância (ou dissimilaridade) entre os *clusters*. Uma distância maior implica que os *clusters* são mais distintos. Neste caso, o COMJUR é muito diferente do CETEM e do COMENTA2 relativamente à distância que estes têm entre si.

Parece-nos, no entanto, que esta diferença é provocada não pelo perfil das unidades linguísticas que nele estão representadas, mas pela diferença do número de ocorrências. Para perceber se esta diferença é causada pela diferença do número de ocorrências, procedemos ao cálculo da frequência relativa (Anexo 4), e executámos um novo *hierarchical clustering*, cujo dendrograma se apresenta no Gráfico 12:

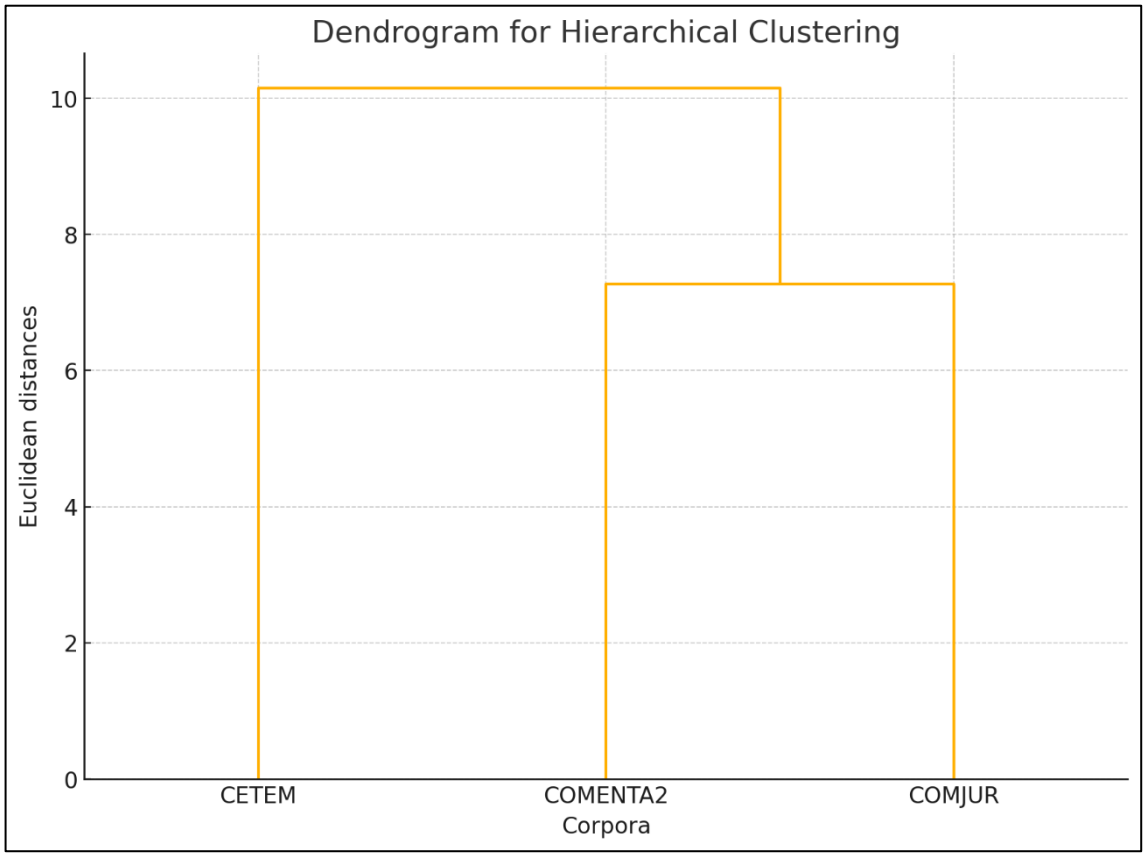


Gráfico 12: Hierarchical clustering com as frequências relativas.

No dendrograma baseado em frequência absoluta (Gráfico 11), CETEM e COMENTA2 agrupam-se primeiro, enquanto COMJUR se une ao grupo posteriormente,

a uma distância muito maior. Isto sugere que, em termos de contagem total de elementos (frequência absoluta), CETEM e COMENTA2 são mais similares entre si, enquanto COMJUR se destaca por apresentar uma contagem absoluta muito diferente dos outros dois *corpora*. No dendrograma baseado na frequência relativa, COMENTA2 e COMJUR agrupam-se primeiro, enquanto o CETEM se une ao grupo depois, a uma distância maior. Esta diferença parece indicar que, em termos de frequência relativa (proporção dos elementos dentro do *corpus*), COMENTA2 e COMJUR partilham mais características em comum, enquanto o CETEM é mais distinto.

Se compararmos as estruturas, observamos que estes dendrogramas indicam que o tipo de medida (relativa ou absoluta) influencia a perceção de similaridade entre os *corpora*. COMJUR tem um comportamento de frequência absoluta diferente, mas proporcionalmente é mais próximo de COMENTA2. Já CETEM e COMENTA2 compartilham uma contagem total mais próxima, mas diferem quando observamos as proporções internas.

Estes dados mostram-nos que embora a análise estatística não revele diferenças significativas entre os *subcorpora*, se observarmos essas diferenças de um ponto de vista praxeológico, podem surgir padrões distintivos. E mostram-nos, também, que embora estas diferenças isoladamente não sejam relevantes, quando analisadas no seu conjunto, são suficientes para agrupar textos "aparentados"[47]. Finalmente, mostram-nos que, para a análise de texto, é possível aplicar técnicas de *data* mining que permitam modelar os dados e descobrir padrões de significado. Se ao longo da nossa análise dos fenómenos microlinguísticos tínhamos dado conta de diferenças entre os *subcorpora*, o *hierarquical cluster* mostra que o conjunto destas diferenças é suficiente para distinguir o CETEM do conjunto formado pelo COMENTA2 e pelo COMJUR.

5.6. Tipos discursivos

A segunda fase do nosso trabalho de análise corresponde à análise dos tipos discursivos presentes no *corpus* COMENTA2, comparando-o com alguns dados do

[47] Relembramos que o COMJUR foi recolhido para este trabalho, utilizando os mesmos critérios de seleção que tinham sido usados para o G&T.Comenta.

CETEM2. Os primeiros dados que podemos observar no Gráfico 13 é a diferença dos TD convocados pelo comentário e pela notícia:

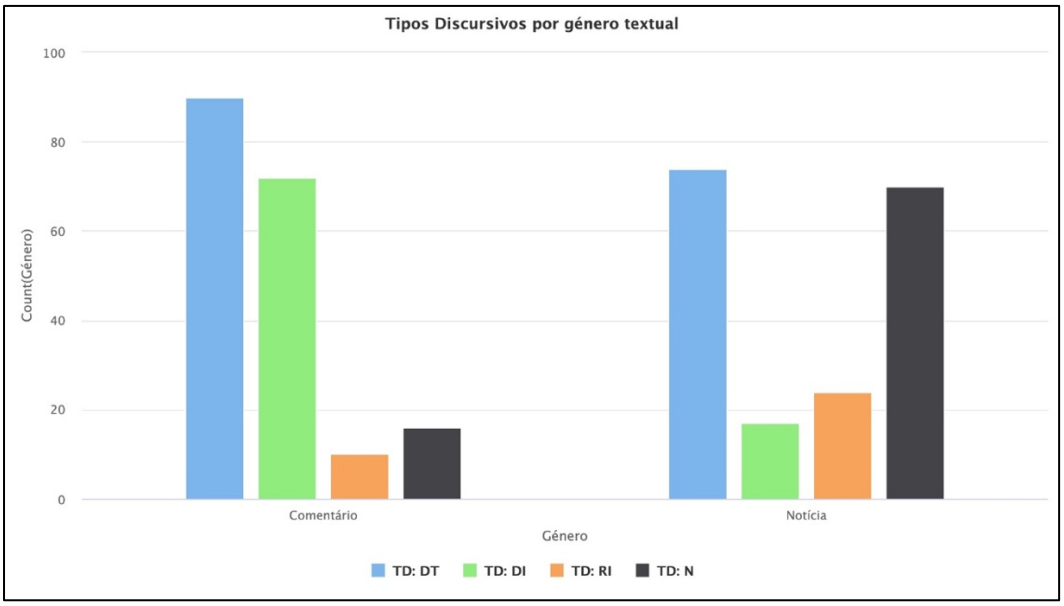


Gráfico 13: Distribuição dos tipos discursivos por género textual.

No comentário destacam-se o discurso teórico e o discurso interativo, e na notícia a narração ocupa o segundo lugar dos TD mais recorrentes, enquanto o DI tem uma presença muito inferior. Também o TD relato interativo tem mais ocorrências na notícia, ocupando o terceiro lugar. Além de podermos observar pelo gráfico que existe uma diferenças nos TD mais frequentes, podemos também constatar que existe maior distribuição dos quatro TD ao longo do *corpus* Notícia. Na secção seguinte vamos analisar com mais pormenor estes dados, recorrendo a uma tabela de frequência e a matrizes de correlação dos tipos discursivos.

5.6.1. Tabela de frequências e matriz de correlação

A tabela de frequências dos TD para o comentário está disponível no Anexo 5, e permite-nos quantificar o número de ocorrências de cada TD para cada texto (coluna ID). Nesta tabela estão incluídas todas as ocorrências, distinguindo os TD que ocorrem em contexto de citação (Cit). O número de ocorrências dos TD em cada texto é muito variável porque, como referimos anteriormente, os textos raramente são constituídos por um único TD. No entanto, podemos ver que o texto 775 é uma exceção e revela marcas de um só tipo, neste caso, o Discurso Teórico. Para darmos conta da relação estatística entre os TD, precisamos de recorrer a uma matriz de correlação.

A matriz de correlação (Tabela 17) mostra que existem pares de tipos discursivos que ocorrem com mais frequência ao longo dos textos analisados, enquanto outros pares de tipos discursivos mostram uma correlação negativa, significando que quando um dos TD ocorre com mais frequência num texto, a probabilidade é que o segundo TD não ocorra. É importante notar que, na leitura da matriz uma correlação forte não implica uma causa direta e que outros fatores podem afetar as variáveis.

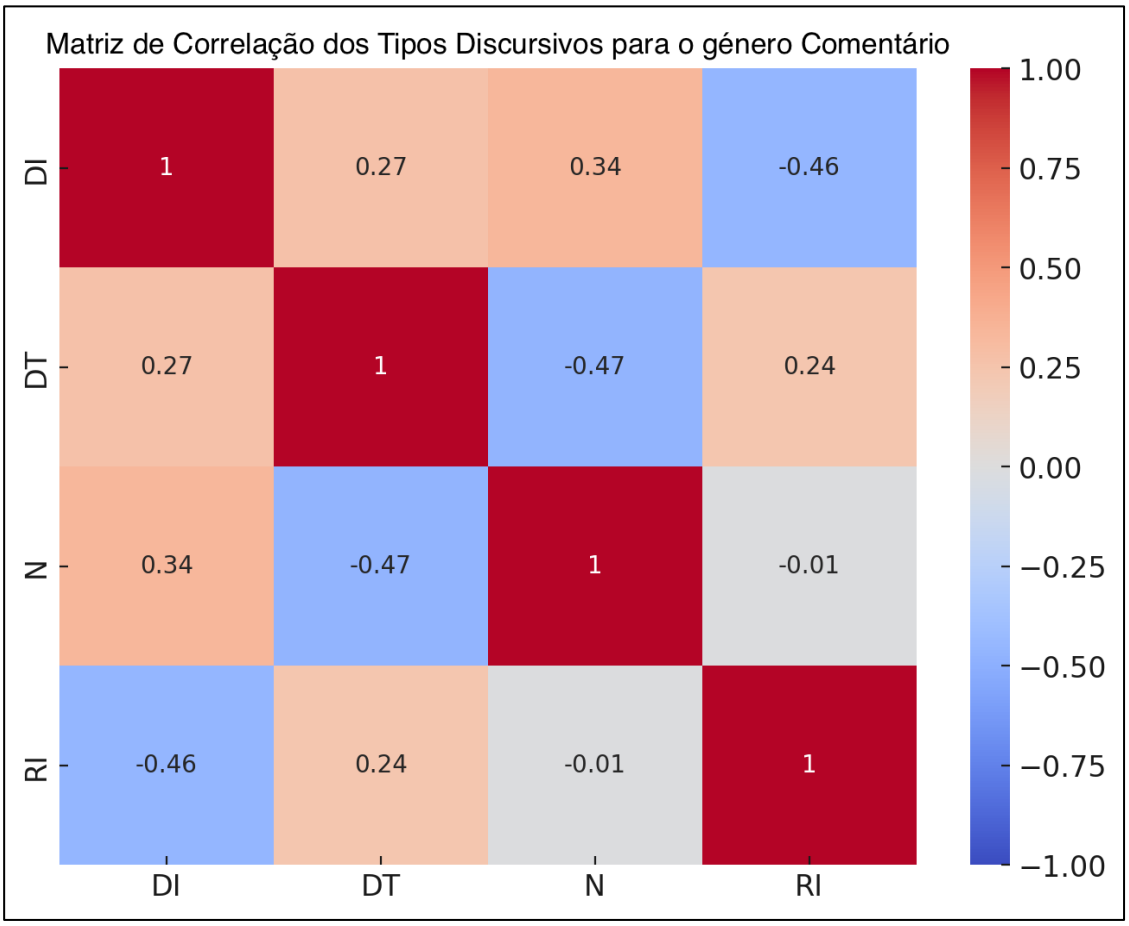


Tabela 15: Matriz de Correlação dos Tipos Discursivos do género Comentário.

A matriz obtida revela algumas correlações:

- **DI e DT**: Correlação positiva de 0.27, indicando uma relação moderada onde textos com muitas ocorrências de DI tendem a ter também mais ocorrências de DT.
- **DI e N**: Correlação positiva de 0.34, sugerindo que textos com mais DI tendem a ter também mais N.
- **DI e RI**: Correlação negativa de -0.46, indicando uma tendência de que quando há mais DI, há menos RI.

- **DT e N**: Correlação negativa de -0.47, sugerindo que quando há mais DT, há tendencialmente menos N.
- **DT e RI**: Correlação positiva de 0.24, mostrando uma leve tendência de que mais DT pode coincidir com mais RI.
- **N e RI**: Correlação muito baixa de -0.01, praticamente indicando que não há relação linear entre N e RI.

  De acordo com o observado, podemos agrupar os TD semelhantes da seguinte forma:
- **DI** e **N** podem ser vistos como relacionados moderadamente positivos (0.34).
- Entre **DI** e **RI** existe uma relação negativa moderada (-0.46).
- Entre **DT** e **N** ocorre uma relação negativa moderada (-0.47).
- As restantes relações são fracas ou nulas, demonstrando pouca influência entre si.

Vamos analisar e comparar com a matriz de correlação do *corpus* CETEM:

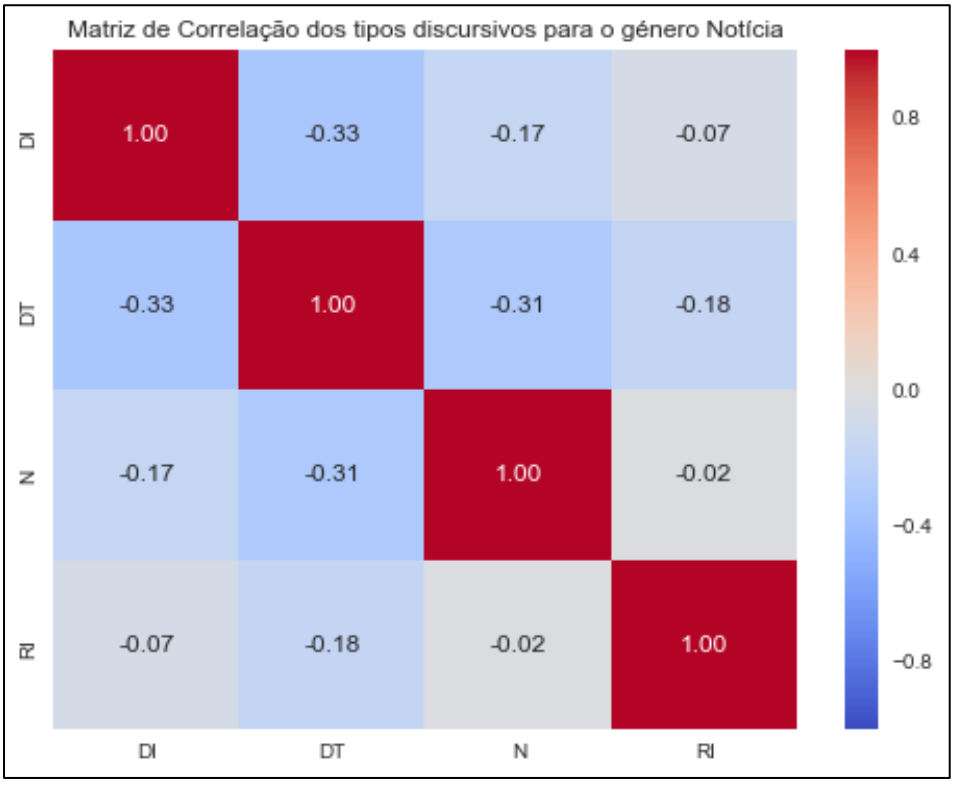


Tabela 16: Matriz de Correlação para o género "Notícia".

A Tabela 18 mostra-nos que a correlação dos tipos discursivos é diferente para o género Notícia:

- Correlação do **DI com**:
    - **DT**: Correlação negativa moderada (-**0.33**). Quando o discurso interativo aumenta, o discurso teórico tende a diminuir.
    - **N**: Correlação negativa fraca (-**0.17**). Há pouca relação entre discurso interativo e narração.
    - **RI**: Correlação próxima de zero (-**0.07**). O discurso interativo e o relato interativo quase não se influenciam.
- Correlação do **DT com**:
    - **DI**: Correlação negativa moderada (**-0.33**) (como mostrado anteriormente).
    - **N**: Correlação negativa moderada (**-0.31**). Quando o discurso teórico ocorre com mais frequência, a narração tende a diminuir.
    - **RI**: Correlação negativa fraca (**-0.18**). Relação fraca e negativa entre discurso teórico e relato interativo.
- Correlação do **N**:
    - **DI**: Correlação negativa fraca (-**0.17**) (como mostrado anteriormente).
    - **DT**: Correlação negativa moderada (-**0.31**).
    - **RI**: Correlação muito fraca ou inexistente (-**0.02**). Não há relação entre narração e relato interativo.
- Correlação do **RI**:
    - DI: Correlação muito fraca (-**0.07**).
    - DT: Correlação negativa fraca (-**0.18**).
    - N: Correlação próxima de zero (-**0.02**).

Significa que todas as correlações entre os TD para o género Notícia são negativas, e que as correlações mais fortes (embora moderadas) são entre os pares DI / DT, e DT / N.

Se compararmos as matrizes, conseguimos observar algumas diferenças importantes:

Par de Tipos	Correlação Notícia	Correlação Comentário	Diferença
DI e DT	-0.33	0.27	0.6
DI e N	-0.17	0.34	0.51
DI e RI	-0.07	-0.46	-0.39
DT e N	-0.31	-0.47	-0.16
DT e RI	-0.18	0.24	0.42
N e RI	-0.02	-0.01	0.01

Tabela 17: Síntese comparativa entre as matrizes de correlação.

De acordo com a Tabela 19, existem diferenças significativas:

- DI e DT: no Comentário há uma correlação positiva (0.27) enquanto na Notícia ocorre uma correlação negativa moderada (-0.33).
- DI e N: no Comentário, DI e N têm uma correlação positiva (0.34) enquanto na Notícia, essa relação é fraca e negativa (-0.17).
- DT e RI: no Comentário têm uma relação positiva (0.24) enquanto na Notícia é uma relação negativa fraca (-0.18).
- DI e RI: no Comentário têm uma relação negativa mais forte (-0.46) enquanto na Notícia é próxima de zero (-0.07).

Esta comparação mostra-nos que os TD, nos segmentos que pertencem ao género Comentário, apresentam correlações mais fortes e positivas, enquanto no género Notícia, as correlações são mais fracas e são todas negativas. No entanto, é necessário observar alguns aspetos quanto medimos esta informação através do método de correlação de Pearson (Schober & Schwarte, 2018): o primeiro aspeto é que o método de correlação de Pearson mede apenas relações lineares entre duas variáveis. Isto significa que, se as relações forem não lineares (exponencial ou curvilínea), o coeficiente pode falhar em captar a relação, resultando em correlações próximas de zero, mesmo que exista algum tipo de relação. Na análise dos nossos dados, podemos afirmar que quanto mais homogéneo for o padrão dos tipos discursivos, maior é a probabilidade do coeficiente captar uma relação linear (positiva ou negativa). Pelo

contrário, se o padrão dos tipos discursivos for mais heterogéneo, maior é a probabilidade de ocorrerem relações não lineares, dificultando a deteção pelo método.

O segundo ponto a tomar em conta, é que este método depende da variabilidade das variáveis em análise. Se uma das variáveis tiver baixa variabilidade (ou for quase constante), a correlação será próxima de zero, mesmo que exista uma relação. O terceiro ponto é que o método de correlação de Pearson é sensível a *outliers* (valores extremos), que podem distorcer as correlações, aumentando ou diminuindo a sua magnitude. O quarto aspeto é que este coeficiente é sensível ao tamanho da amostra, o que significa que conjuntos pequenos de dados podem produzir correlações instáveis, influenciadas por flutuações aleatórias. O coeficiente de Pearson pressupõe ainda que os dados sejam quantitativos e contínuos, mas para a análise de texto, as sequências dos TD têm de ser codificadas como frequências ou proporções, o que também pode influenciar os resultados. Todos estes fatores resumem-se ao facto de que quanto mais equilibradas e previsíveis forem as distribuições dos TD, mais lineares são as correlações, enquanto se, pelo contrário, a distribuição for mais dispersa ou desigual, maior é a deteção de padrões frequentes.

Finalmente, importa referir o sexto e último ponto da análise que é a interpretação estatística que deve ser feita: o coeficiente de Pearson mede a associação e não a causalidade, o que significa que as diferenças nas correlações entre os *corpora* não indicam necessariamente que os géneros textuais "causam" os padrões observados.

Os resultados obtidos na Tabela 17 mostram que existem correlações positivas e negativas entre os diferentes TD, significando que, através da matriz de correlações é possível observar o "comportamento" dos TD no *corpus*. E que este comportamento é específico deste *corpus*, como podemos constatar pela comparação com a matriz de correlação do *corpus* CETEM. Não só os TD parecem excluir ou condicionar a ausência de determinados TD como existem géneros (no nosso caso o género Comentário) que revelam uma dependência positiva entre os TD. A nossa hipótese sobre estes dados é que esta diferença entre relação positiva e relação negativa entre os TD depende da homogeneidade ou heterogeneidade dos TD que ocorrem nos textos em análise. Tal como vimos no primeiro capítulo deste trabalho, os TD articulam-se com os géneros em

vários níveis sendo que no segundo nível, os géneros estabilizam e mobilizam um ou mais TD e que o mesmo TD pode apresentar especificidades em diferentes géneros.

Se os tipos discursivos, como observa Coutinho (2019), são segmentos com características linguísticas estáveis, que constituem um conjunto limitado de possibilidades, consideramos que, embora o comentário e a notícia mobilizem TD diferentes, a relação que os TD mobilizados estabelecem entre si é também diferente: o género comentário mobiliza maioritariamente os tipos discursivos DT e DI, e os textos pertencentes ao género notícia mobilizam mais DT e N, mas a correlação que o DT estabelece com DI e N não é igual. Considerando que a matriz de correlação é uma fórmula estatística que não tem acesso a outros dados, a única explicação que se coloca para esta diferença é que a distribuição dos tipos discursivos é também diferente: no género notícia, os TD são mais heterogéneos do que no comentário. A leitura destes resultados deve ser feita com algumas cautelas porque embora mostre diferenças na forma como os TD se relacionam entre si, e ofereça pistas de análise, aquilo que observamos é uma relação de associação e não de causalidade. Isto significa que a explicação para as diferenças de correlação entre os *corpora* deve ter em conta as propriedades do coeficiente de Pearson que foi utlizado. De acordo com estas propriedades, as correlações não refletem apenas diferenças no modo como os TD se relacionam entre si, mas também a forma como os dados são estruturados e analisados estatisticamente. Partindo desta reserva acerca do coeficiente de Pearson, a leitura das matrizes de correlação, sustentadas pelo Gráfico 13, parece ser que quanto mais heterogénea é a distribuição dos TD mais fracas são as correlações, tal como observamos no género notícia, enquanto no género comentário, que tem uma distribuição mais homogénea, as correlações são tendencialmente mais fortes e positivas. Assumimos que se trata de uma hipótese que necessita de mais dados, mas que, confirmando-se, revela que uma matriz de correlação oferece uma nova forma de analisar como os TD se distribuem ao longo dos textos.

A análise das matrizes de correlação sugere ainda que, de acordo com o conceito de parâmetros de género, o género Comentário tem uma distribuição mais homogénea dos TD, refletindo uma maior estabilidade e organização interna, sugerindo que os tipos

discursivos e o padrão como eles se correlacionam ao longo do texto constituem-se como um parâmetro de género, visualizável através de uma matriz de correlação.

Este facto fez-nos questionar se a forma como os TD surgem no *corpus* influenciam esta correlação, que vamos analisar na secção seguinte.

5.6.2. Sequência dos Tipos Discursivos

A sequência dos TD ao longo dos textos do *corpus* pode ajudar-nos a perceber como os TD interagem entre si e observar se as relações entre os TD estão associadas a padrões contextuais ou estruturais. O Gráfico 14 mostra a posição de cada tipo discursivo ao longo de todo o *corpus*. Cada ponto corresponde a um TD (eixo X) identificado nos textos, e deve ser lido da esquerda para a direita, em que o ponto 0 corresponde ao primeiro TD identificado no *corpus* (no primeiro texto) e o último ponto corresponde à última sequência de TD identificada no último texto do *corpus*. Nele podemos observar que há uma alternância frequente entre DI e DT, o que sugere que estes TD ocorrem frequentemente em conjunto ou em sequência. Os RI e N também aparecem em diversas partes do gráfico, embora com menos frequência em comparação com DI e DT. Outro aspeto observável é a ocorrência de agrupamentos de pontos em certas regiões do gráfico, especialmente em torno de DT e DI, e que pode indicar que esses tipos de TD ocorrem em sequências próximas ou são repetidos consecutivamente.

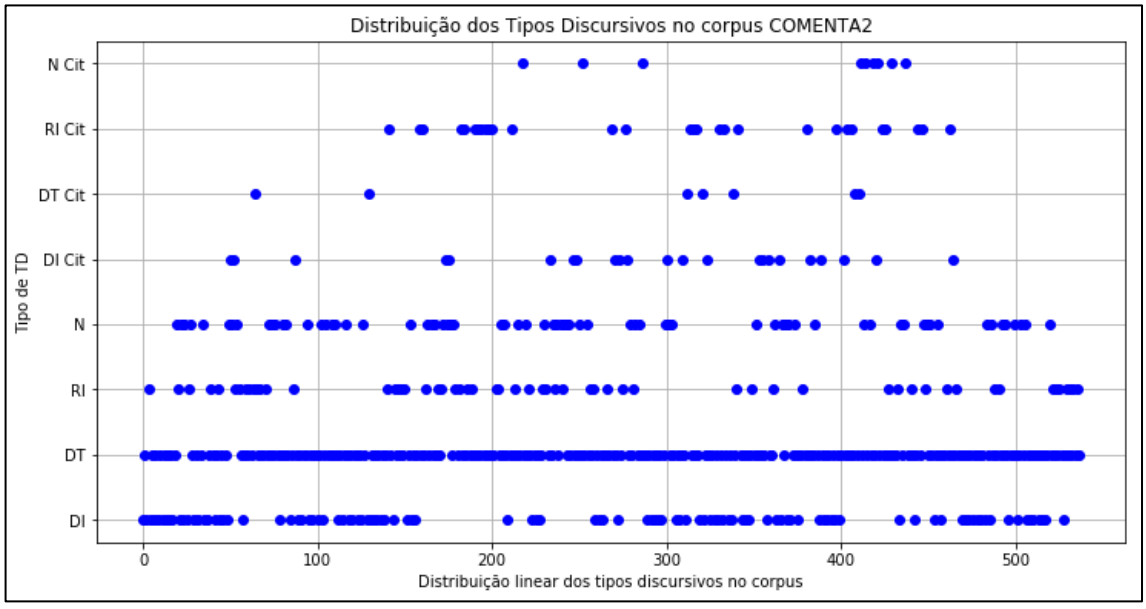


Gráfico 14: Distribuição dos tipos discursivos no corpus.

O gráfico sugere ainda que DI e DT são os tipos discursivos mais dominantes, com uma presença consistente e densa ao longo do *corpus*. Outros tipos, como N Cit e RI Cit (que ocorrem em contexto de citação), surgem menos frequentemente e de maneira mais esparsa. A alternância entre certos tipos discursivos, especialmente DI e DT, pode indicar padrões de uso no *corpus*, enquanto a continuidade de certos tipos discursivos por longos períodos pode indicar partes do *corpus* onde um tipo discursivo específico domina o texto. A análise da sequência e das subsequências dos tipos discursivos mostra que há correlações fortes entre os TD presentes no comentário, e que este atributo tem variações que parecem ser específicas do *corpus* em análise. Sendo uma particularidade do *corpus*, mostra-se fidedigna como variável na classificação do género comentário e, ao analisarmos a sequência dos TD ao longo dos textos, podemos observar que existem padrões associados.

5.6.3. Subsequências dos Tipos Discursivos

Como podemos observar a partir da matriz de correlação (Tabela 17) e do gráfico de sequências (Gráfico 14) há um par de TD dominante ao longo do *corpus*. Esta observação fez-nos questionar se existiriam outras sequências que se repetissem com pares de TD, e o resultado está expresso no Gráfico 15.

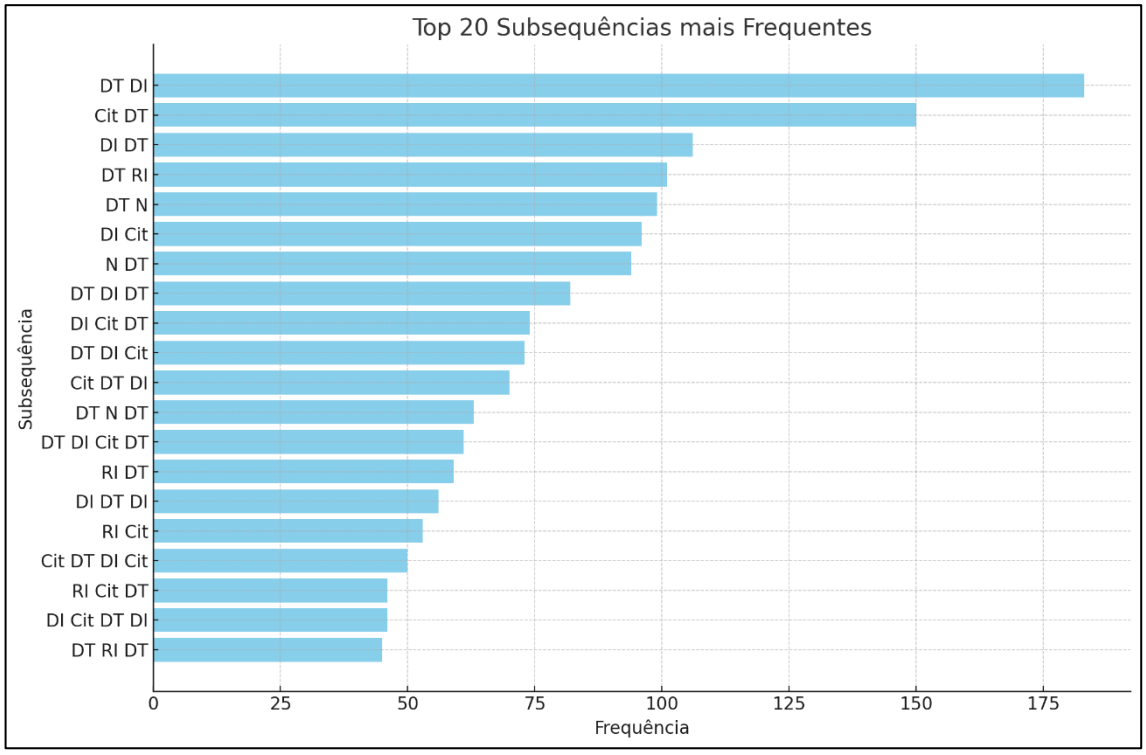


Gráfico 15: Subsequências mais frequentes no corpus.

O Gráfico 15, além de nos mostrar que subsequências são mais comuns no *corpus*, mostra-nos também o comprimento dessas subsequências, sendo que o máximo que encontrámos foi de quatro TD. Embora não possamos afirmar que a diferença entre o comprimento das subsequências seja muito relevante, podemos dizer que quanto maior for o comprimento, mais incomum é a subsequência.

A análise dessas subsequências permite identificar padrões recorrentes na estrutura dos textos e da relação entre diferentes tipos de discurso:

(ii) Há um domínio do Discurso Teórico: está presente em quase todas as subsequências mais frequentes, sendo predominante na maioria dos textos. Esta frequência elevada sugere que o DT desempenha um papel central na composição dos textos analisados. As subsequências mais frequentes começam ou contêm o DT combinado com outros tipos discursivos, como o DI e o RI.

(ii) Relação frequente entre DT e DI: a subsequência mais frequente é DT DI, seguida por Cit DT e DI DT, evidenciando que há uma transição frequente entre o Discurso Teórico e o Discurso Interativo.

(iii) Presença da Citação: a presença de Citação (Cit) em várias subsequências, como Cit DT e DI Cit, mostra que a utilização de citações desempenha um papel importante na construção dos textos. A frequência da subsequência Cit DT sugere que as citações são frequentemente seguidas por segmentos de natureza teórica ou explicativa.

(iv) Variedade nas combinações dos tipos discursivos: o gráfico evidencia uma diversidade de combinações entre os tipos discursivos, o que sugere que os textos apresentam uma estrutura heterogénea, com alternância entre diferentes tipos de discurso, onde tipos discursivos secundários desempenham papéis complementares.

(v) Subsequências longas e complexas: além de subsequências curtas e frequentes, como DT DI e DI DT, também observamos subsequências mais longas e complexas, como DT DI Cit DT e DT RI Cit DT. A presença dessas combinações indica que os textos têm uma organização interna mais elaborada.

O gráfico mostra que o DT é um tipo discursivo fundamental na estrutura dos textos, frequentemente combinado com outros tipos discursivos, como o DI e o RI. A presença frequente de citações sugere que a intertextualidade é uma característica marcante dos textos, reforçando a natureza argumentativa ou explicativa das instâncias analisadas. A análise das subsequências mais frequentes fornece uma visão detalhada da estrutura interna dos textos.

Se compararmos estes dados com os dados do Gráfico 13, podemos afirmar que a alternância dos TD ao longo do *corpus* não é aleatória: há uma preferência por subsequências curtas, que indicam TD dominantes, com uma repetição do padrão: o DT, por exemplo, repete-se nas cinco primeiras subsequências, desempenhando, por isso, um papel mais constante ao longo do *corpus*, enquanto RI e N têm uma distribuição mais isolada.

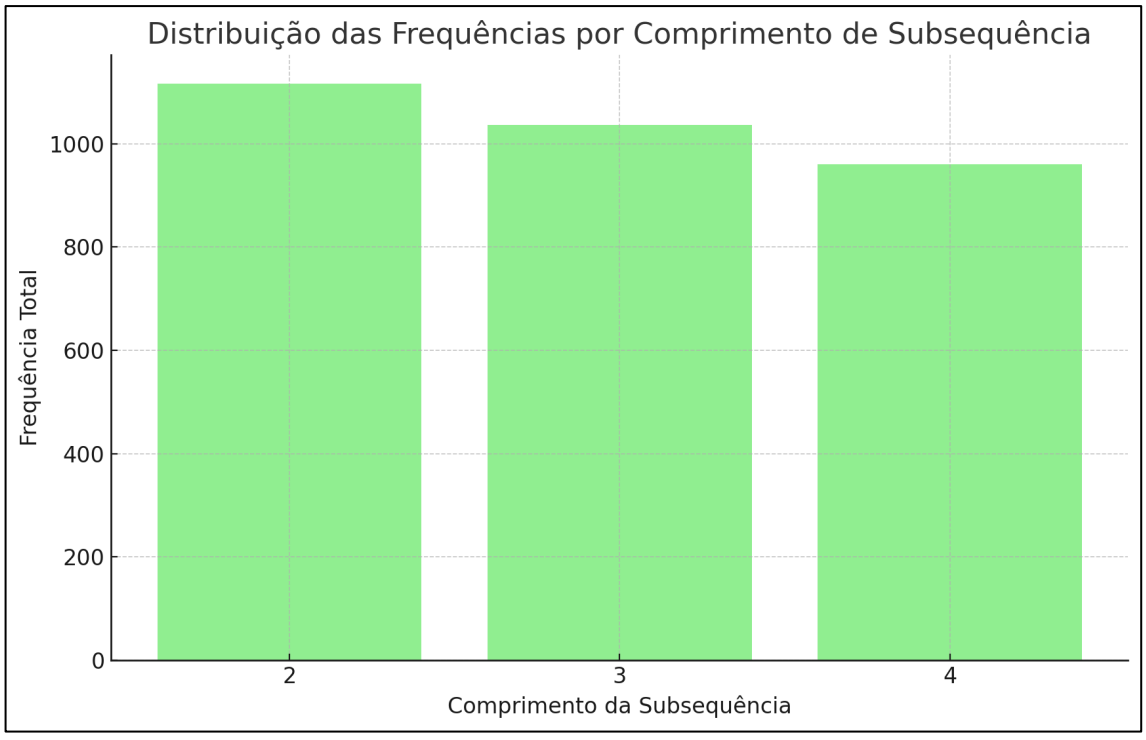


Gráfico 16: Relação das frequências por comprimento de subsequência.

Ao observamos os Gráficos 15 e 16 notamos que existe uma estabilização das subsequências mais curtas (duas sequências), enquanto as subsequências mais longas (>2) são mais irregulares ao longo do *corpus*. Esta estabilização parece relevar da forma como os TD se relacionam com o plano de texto. Bronckart destaca o papel do plano de texto na organização do conteúdo temático, mas também afirma que o

plano é determinado pela combinação dos tipos discursivos, das sequências e de outras formas de organização presentes no texto (Bronckart, 1997: 122), das quais já abordámos os mecanismos de coesão verbal e nominal. E sublinha também (Bronckart, 2008: 52) que a análise do plano de texto envolve a identificação de secções e partes de um texto que podem ser delimitadas por critérios de expressão e conteúdo, e que esta análise é fundamental para entender como os diferentes tipos discursivos se organizam e se manifestam dentro de um género textual específico. Dada a relevância que a organização dos TD tem para o plano do texto, a análise da sequencialidade e do tipo de sequências de TD que ocorrem no *corpus* COMENTA2 mostra padrões que contribuem para a caracterização do género comentário.

Ao observar os dados obtidos, podemos perceber como os padrões identificados nas subsequências textuais podem funcionar como marcadores de género. A estabilização das subsequências mais curtas, por exemplo, pode indicar a presença de mecanismos linguísticos que são típicos de um género específico, como o género comentário. Esses padrões revelam a repetição de certos tipos de marcadores linguísticos (como pronomes, advérbios ou locuções temporais), que, quando analisados, fornecem pistas para a identificação do género. A regularidade dos padrões nas subsequências mais curtas reflete, também a aplicação consistente dos mecanismos de realização textual, que são responsáveis pela manutenção da identidade do género. Ao integrar as ferramentas de análise quantitativa, como o *text mining* e o *data mining*, conseguimos não apenas identificar essas marcas linguísticas, mas também validar a ideia de que os padrões de TD podem atuar como marcadores de género, ajudando a delimitar as fronteiras e as características específicas do género comentário.

Em resumo:

A caracterização dos tipos linguísticos dos *corpora*, efetuada na primeira fase, mostra-nos que não existem diferenças estatisticamente significativas que permitam estabelecer diferenças claras com as variantes analisadas. No entanto, embora os testes de relevância estatística não revelem diferenças nas suas médias, as diferenças que existem são relevantes.

Vejamos os dados obtidos dos pronomes e dos advérbios. Embora não haja diferenças estatísticas relevantes, o número de ocorrências mostra que o COMENTA2 tem uma predominância de pronomes e advérbios de lugar que podem ter uma interpretação deítica, e a predominância dos pronomes de primeira e segunda pessoa, juntamente com o predomínio dos advérbios, parecem indiciar uma convergência entre a localização espacial e o ponto de vista do enunciador.

Por outro lado, o elevado número de ocorrências de advérbios e locuções temporais no *subcorpus* CETEM está ligado à própria natureza da atividade em que é produzido texto (atividade jornalística) e constitui um traço específico do género notícia (Miranda, 2010). O *corpus* COMENTA2, apesar de incluir textos pertencentes à atividade jornalística, não tem um número relevante de advérbios e locuções temporais, facto que pode indiciar um traço intrínseco à prática do comentário.

Outro aspeto que se evidencia da análise quantitativa dos dados, na primeira fase, é que as variáveis, quando analisadas isoladamente, não têm relevância estatística entre os diferentes *subcorpora*. No entanto, quando as variáveis são analisadas em conjunto, como mostramos nos Gráficos 11 e 12, existem diferenças que distinguem o conjunto formado pelos *subcorpora* COMJUR e COMENTA2 comparativamente ao *subcorpus* CETEM. Isto parece mostrar que, embora a análise quantitativa dos tipos linguísticos não mostre diferenças, a sua coocorrência permite agrupar determinados conjuntos de textos. O que nos leva a concluir que uma análise do tipo micro não é suficiente para encontrar traços distintivos entre os diferentes textos. Os testes de significância estatística mostram que uma análise quantitativa que trabalhe exclusivamente com médias estatísticas anula diferenças que, no seu conjunto, podem ser relevantes. Significa, também, que quando lidamos com grandes volumes de dados, o tipo de análise que é feita tem implicações nos resultados que obtemos: usando apenas médias, as diferenças que, em conjunto podem ser interpretadas qualitativamente, desaparecem, tal como observou McEnery & Wilson (1996: 75-77). Isto parece mostrar que o uso de ferramentas automáticas de texto, baseado em médias estatísticas pode mascarar diferenças significativas, evidenciando a importância de uma análise mais qualitativa e holística. Esta hipótese é reforçada pelos resultados dos dendrogramas obtidos através do *hierarquical cluster* (Gráficos 11 e 12) que realizámos.

Estes dendrogramas mostram também como a referência de medida utilizado altera a perceção dos resultados: se a frequência absoluta mostra semelhanças entre CETEM e COMENTA, com afastamento do COMJUR, a frequência relativa mostra que os dados obtidos na análise agrupam o COMENTA e o COMJUR. A nossa hipótese para a interpretação destes resultados é que existem diferenças ao nível das unidades linguísticas que permitem separar o comentário de outras práticas textuais (neste caso, a notícia). A comparação dos resultados dos dendrogramas mostra-nos que quando analisamos texto, e não fenómenos microlinguísticos isolados, os dados têm de ser modelados para que possam ser analisados em conjunto. E mostra-nos, também, que é possível aplicar técnicas de *data mining* para analisar textos, e que os padrões que estas técnicas revelam são mais importantes que a contagem de ocorrências:

> *Simple frequency tables of linguistic features in texts and corpora can often mask more general patterns of similarity and difference which may help to explain better why particular features, varieties and languages behave in the way that they do: multivariate statistical techniques can help the corpus linguist to extract these hidden patterns from the raw frequency data*. (McEnery, 2001:91).

Na análise das subsequências de tipos discursivos, o COMENTA2 caracteriza-se por subsequências mais curtas e repetitivas, o que pode estar associado à construção de sequências discursivas que operam de forma mais homogénea, enquanto as subsequências mais longas podem refletir estruturas mais complexas.

A análise dos tipos discursivos e as unidades linguísticas no género comentário revelam padrões que podem ser explicados pelos parâmetros e marcadores de género. Os parâmetros de género, que se referem aos elementos previsíveis e recorrentes, que organizam um género textual, funcionam como modelos gerais para a produção e interpretação do texto. No caso do género comentário, a regularidade observada nos tipos discursivos no *subcorpus* COMENTA2, caracterizado por subsequências curtas e repetitivas, sugere uma organização interna mais estável e homogénea, que pode ser interpretado como um parâmetro de género. Por outro lado, os marcadores de género, entendidos como os mecanismos semióticos (verbais e não verbais), atuam como "pistas" para a identificação desses parâmetros. Na nossa análise, os marcadores de género podem ser detetados na forma como as unidades linguísticas, como pronomes,

advérbios de lugar e locuções temporais são realizados nos textos. No caso específico do COMENTA2, a predominância de pronomes de primeira e segunda pessoa, assim como o uso de advérbios de lugar, são identificados como marcadores de género.

6. Análise do modelo

Na terceira e última fase do nosso trabalho, construímos um modelo de classificação que permite, através de variáveis meso (TD) e macro (tema e atividade), classificar os segmentos de texto por género (Notícia ou Comentário). Nas seções seguintes vamos analisar as métricas obtidas, começando pelo desempenho do modelo, e relacioná-las com os dados obtidos.

6.1. Avaliação do modelo

A nossa avaliação começa por analisar a *Confusion Matrix* [48](Figura 20), uma vez que é a partir dos valores desta tabela que obtemos outras métricas de desempenho.

Confusion Matrix

	true Notícia	true Comentário	class precision
pred. Notícia	50	6	89.29%
pred. Comentário	3	47	94.00%
class recall	94.34%	88.68%	

Figura 19: Confusion matrix do Modelo de GLM.

Podemos observar na tabela que a Previsão "Notícia" obteve 50 casos que foram corretamente classificados como *True* "Notícia", enquanto em *True*"Comentário", 6 casos foram incorretamente classificados como "Notícia" (falsos positivos para a classe "Notícia"). A *Precision* para "Notícia" é:

$$\frac{50}{(50+6)} = \approx 89{,}29\%.$$

A Previsão "Comentário" obteve 3 casos que foram incorretamente classificados como *True* "Notícia" (falsos negativos para a classe "Notícia") e para *True* "Comentário",

[48] A *confusion matrix* é uma tabela onde são registadas as contagens dos registos de teste correta e incorretamente previstos pelo modelo. É um dos elementos de avaliação do desempenho do modelo (Tan et al., 2018: 150).

47 casos que foram corretamente classificados como “Comentário”. A *Precision*[49] para “Comentário” é:

$$\frac{47}{(47+3)} = \approx 94{,}00\%.$$

Partindo destes valores, obtemos também algumas métricas derivadas como a *Recall*[50]:

- ***Recall* da classe “Notícia”**: $\frac{50}{(50+3)} = \approx$ **94,34%**, representando a capacidade de identificar corretamente instâncias de “Notícia”.
- ***Recall* da classe “Comentário”**: $\frac{47}{(47+6)} = \approx$ **88,68%**, refletindo a capacidade de identificar corretamente “Comentário”.

Na Figura 21, podemos observar as restantes métricas de desempenho do modelo GLM, que derivam da *confusion matrix,* e que passamos a analisar:

Generalized Linear Model – Performance

Performances

Criterion	Value	Standard Deviation
Accuracy	91.5%	± 2.1%
Classification Error	8.5%	± 2.1%
AUC	94.6%	± 5.0%
Precision	94.4%	± 5.2%
Recall	88.9%	± 7.6%
F Measure	91.2%	± 2.3%
Sensitivity	88.9%	± 7.6%
Specificity	94.4%	± 5.2%

Figura 20: Métricas de desempenho do Modelo de GLM.

1. **Accuracy (91,5% ± 2,1%)**: É a proporção total de previsões corretas. O modelo classifica corretamente aproximadamente 91,5% das instâncias.

[49] A ***Precision*** mede a proporção de exemplos que o modelo classificou como positivos que realmente são positivos, e é útil quando queremos minimizar **falsos positivos**. É calculada usando a seguinte fórmula: $Precision = \frac{Verdadeiros\ Positivos\ (True\ Positive)}{Verdadeiros\ Positivos\ (True\ Positive)+Falsos\ Positivos\ (FP)}$

[50] A ***Recall*** mede a proporção de exemplos realmente positivos que o modelo identificou corretamente, e é útil quando queremos minimizar **falsos negativos**. É calculada usando a seguinte fórmula: $Recall = \frac{Verdadeiros\ Positivos\ (True\ Positive)}{Verdadeiros\ Positivos\ (True\ Positive)+Falsos\ Negativos\ (FN)}$

2. **Classification Error (8,5% ± 2,1%)**: Sendo uma medida complementar à *Accuracy*, refere-se à proporção de previsões incorretas. Apenas 8,5% das previsões foram classificadas erradamente.

3. **AUC** [51]**(94,6% ± 5,0%)**: Área sob a curva ROC. Um AUC de 94,6% indica que o modelo tem um excelente desempenho na separação entre classes positivas e negativas.

4. **Precision (Precisão) (94,4% ± 5,2%)**: Como já referimos, a Precisão mostra que, das predições classificadas como positivas, 94,4% eram realmente positivas.

5. **Recall/Sensitivity (88,9% ± 7,6%)**: Dos casos positivos reais, 88,9% foram corretamente identificados pelo modelo.

6. **F-Measure (91,2% ± 2,3%)**: Combina precisão e *recall* numa única métrica, equilibrando os dois aspetos. O valor obtido mostra que o modelo é consistente.

7. **Specificity (94,4% ± 5,2%)**: Mede a proporção de verdadeiros negativos corretamente identificados. Os valores mostram que o modelo evita com facilidade os falsos positivos.

Os dados obtidos mostram que as margens de erro (*Standard Deviation*) são relativamente pequenas e sugerem que o modelo é robusto e estável. Há equilíbrio entre sensibilidade (*sensitivity*) (88.9%) e especificidade (*specificity*) (94.4%), indicando que o modelo tem um desempenho sólido para ambas, e apresenta equilíbrio entre as classes, como podemos ver nos valores de *Precision* e *Recall* semelhantes para "Notícia" e "Comentário" obtidos na *Confusion Matrix*. Finalmente, o baixo valor de falsos positivos em ambas a classe reflete a capacidade do modelo de identificar corretamente os verdadeiros negativos, revelado pelo valor da especificidade (94,4%).

O modelo mostra um bom desempenho, com uma boa *Precision* e *Recall*. As métricas derivadas da *Confusion Matrix* corroboram as métricas gerais do modelo

[51] AUC é uma métrica que mede a capacidade do modelo em distinguir entre classes positivas e negativas, e é particularmente útil quando há dados desequilibrados.

apresentadas nas Figura 20 e 21, e sugerem que o modelo obtido é confiável e equilibrado para classificar tanto "Notícia" quanto "Comentário".

6.2. Pesos, Correlações e Previsões

Como referimos anteriormente, o GLM atribui pesos às diferentes classes. Na Tabela 20 observamos o peso que foi atribuído a cada uma das classes. Embora estes valores sejam atribuídos pelo próprio algoritmo[52] durante a *Cross-validation*, o *RapidMiner* permite ao utilizador alterar estes valores para otimizar os resultados. Esta opção é particularmente importante quando existe um grande desequilíbrio nos valores dos atributos. Por exemplo, se houver um atributo excessivamente representado e se o algoritmo tiver dúvidas na previsão, a probabilidade é que escolha o atributo sobrerepresentado. Neste trabalho, optámos por aceitar os pesos atribuídos pelo GLM uma vez que os valores são ajustados automaticamente pelo algoritmo durante o processo de treino, de modo a minimizar a margem de erro e a maximizar a precisão do modelo. No entanto, existem alguns cenários onde poderíamos alterar os pesos atribuídos:

1. Interpretação ou ajustes manuais: quando o conhecimento sobre os dados nos permitem ajuizar sobre o peso indevido atribuído a uma determinada classe (Chein, 2019).

2. Para combater o *overfitting* dos dados: quando o modelo está demasiadamente ajustado aos dados, problema que pode ser identificado quando há um peso excessivamente grande atribuído a uma determinada classe (Witten & Frank, 2005).

3. Problemas com variáveis correlacionáveis: quando existem duas variáveis fortemente correlacionadas (Dormann et al., 2013).

A alteração dos parâmetros do modelo, além da perda de otimização feita pelo modelo, implicaria a revisão e reavaliação do desempenho e uma reanálise dos resultados obtidos com a alteração, tendo em conta os ajustes efetuados. Tendo o

[52] Os valores são calculados através de um processo chamado Máxima Verossemelhança (*Maximum likelihood*) (Tan et al., 2018).

modelo obtido resultados satisfatórios, não nos parece oportuno, neste momento, proceder a alterações, mas no final do capítulo, serão dadas algumas orientações com vista à sua otimização, partindo da leitura dos resultados obtidos.

Neste trabalho, optámos por aceitar os pesos atribuídos pelo GLM porque não temos, neste momento, uma métrica que permita aferir a importância de cada classe.

Attribute	Weight
Atividade	0,313
TD	0,205
Tema	0,163

Tabela 18: Peso por atributo.

De acordo com a tabela, a Atividade tem um peso superior na deteção da variável alvo que é a identificação do género, embora o TD e o Tema tenham valores relativamente próximos:

a) **Atividade (Peso: 0,313)**: É o atributo com o maior peso, indicando que tem uma maior relevância na previsão da variável-alvo (Género). Foi o atributo que mais significativamente contribuiu para o modelo, sendo o principal fator para as variações observadas na variável em análise.

b) **TD (Peso: 0,205**): tem um peso intermédio, o que mostra que também é importante, mas menos do que "Atividade". A relevância ainda é significativa, indicando que há uma correlação considerável com a variável-alvo.

c) **Tema (Peso: 0,163)**: este atributo tem o menor peso entre os três, sugerindo que a sua influência na previsão da variável-alvo é mais limitada. No entanto, a sua contribuição deve ser observada como complementar aos outros dois atributos na análise.

O que os valores na Tabela 20 nos mostram é que a Atividade deve ser prioritária na interpretação dos resultados ou no caso de ser necessário, futuramente, ajustar o modelo, uma vez que é aproximadamente 1,5 vezes mais relevante do que TD e quase 2 vezes mais relevante do que o Tema. Este resultado alinha-se com duas questões

importantes no que toca a noção de género: a primeira relaciona o género com as atividades da linguagem, ou seja, a forma como as atividades comunicativas influenciam a escolha do género; e a segunda concerne a estabilidade do género e a forma como a produção textual evidencia os elementos linguísticos que relevam do género convocado. Se a primeira questão reenvia à relação entre atividades de linguagem e género, a segunda relaciona-se com a estabilidade do género. Como foi destacado por Adam (1997: 670) e Rastier (2001: 299), o género não se limita a um conceito exclusivamente linguístico, sendo, na verdade, um fenómeno social com uma forte dimensão cultural e histórica. Além disso, o género serve como uma referência na textualização, orientando a organização e interpretação dos textos. A natureza dos géneros é essencialmente socio-comunicativa, já que estes funcionam como dispositivos flexíveis e dinâmicos que se transformam com o tempo, o espaço e a maneira como são utilizados pelos agentes textuais. Os géneros emergem das necessidades comunicativas e das atividades sociais, sendo o fundamento da morfogénese genealógica ou da evolução dos géneros. Trata-se, portanto, na validação de uma abordagem descendente, na esteira de Volochinov, que aprofundou a ideia de que os géneros influenciam tanto a organização textual como a estrutura linguística. Volochinov destacou que "cada tipo de comunicação social […] organiza, define e constrói, de maneira única, a forma gramatical e estilística da enunciação, assim como a estrutura do tipo a que pertence: a isso chamamos género" (Volochinov, 1977: 289-290) Essa perspetiva, que coloca os géneros como determinantes na organização linguística, retoma a abordagem metodológica descendente, que propõe o estudo inicial das atividades sociais, seguido da análise dos géneros ou "atos de fala", para, finalmente, analisar as "estruturas linguísticas" (Volochinov, 1977:137-139).

Outra tabela que contém métricas importantes e que corroboram a importância da atividade para o género é a matriz de correlações (Anexo 7). A matriz de correlações apresentada permite avaliar a relação entre as variáveis preditoras — Tema, Atividade e Tipos Discursivos — e a variável alvo Género. As correlações, medidas através do coeficiente de correlação de Pearson, indicam a força e a direção da relação linear entre duas variáveis:

(i) Relação entre Atividade e Género: A variável Atividade apresenta uma correlação significativa com o Género dos textos, sugerindo que o tipo de atividade em que o texto é produzido exerce uma influência considerável na classificação do género textual. De acordo com a matriz de correlações, observa-se que determinadas atividades apresentam uma correlação mais forte com a variável alvo, o que indica que certos géneros de texto estão predominantemente associados a contextos específicos de produção. Esta relação sugere que a Atividade é uma variável relevante na discriminação do género textual, contribuindo para a correta classificação dos textos pelo modelo. A presença de uma correlação moderada a alta entre as variáveis implica que a informação relativa ao contexto de produção não é redundante e acrescenta valor na previsão do género dos textos. Revela, também como as variáveis preditoras se relacionam de maneira significativa com o género textual: a Atividade é uma das variáveis mais influentes na classificação do género, pois reflete o contexto em que o texto é produzido, alinhando-se com o conceito de parâmetros de género. Ou seja, a escolha do género está diretamente ligada ao tipo de atividade social e comunicativa em que o texto se insere, como observado pela correlação significativa entre Atividade e Género.

(ii) Relação entre Tema e Género: o Tema dos textos também apresenta uma correlação relevante com o Género, embora, em geral, os coeficientes sejam ligeiramente inferiores aos observados para a variável Atividade. Esta diferença sugere que, apesar de o tema ser uma variável importante na classificação do género, a sua contribuição é menor em relação à variável Atividade. Determinados temas apresentam correlações mais altas, o que indica que a presença de conteúdos relacionados com áreas específicas (como política, economia ou desporto) pode ser um fator diferenciador para o género textual. No entanto, o facto de os coeficientes serem moderados sugere que o tema, isoladamente, não é suficiente para uma classificação precisa, mas sim quando combinado com outras variáveis, como a Atividade.

(iii) Relação entre Tipos Discursivos e Género: os TD apresentam correlações distintas com o Género. As correlações moderadas observadas entre alguns tipos discursivos, especialmente o DT, e a variável alvo indicam que o tipo de discurso predominante no texto desempenha um papel importante na classificação do género. O DT, por exemplo, apresenta uma correlação positiva mais elevada, sugerindo que este

tipo de discurso está frequentemente associado a textos do género "Notícia". Em contraste, o RI e o Discurso Interativo DI, apresentam correlações mais baixas, o que indica que estes tipos discursivos não contribuem significativamente para a identificação do género do texto. A presença de correlações baixas pode sugerir que esses tipos discursivos ocorrem de forma mais distribuída entre os diferentes géneros, não sendo tão determinantes quanto o DT.

A matriz também permite identificar as correlações entre as variáveis preditoras, o que é relevante para compreender a interação entre elas na classificação do género textual. Observa-se que há correlações moderadas entre Atividade e Tema, o que sugere que determinadas atividades estão associadas a temas específicos. A ausência de correlações muito fortes entre as variáveis preditoras indica que elas fornecem informações complementares ao modelo, sem redundância excessiva. Isto é importante, pois sugere que todas as variáveis contribuem de forma independente para a classificação do género, aumentando a robustez do modelo.

A análise da matriz de correlações evidencia que a Atividade é a variável que apresenta maior influência na classificação do Género dos textos, seguida pelos TD e pelo Tema. A Atividade parece ser um forte indicador do contexto de produção do texto, o que, por sua vez, está altamente relacionado com o género textual. Os Tipos Discursivos também oferecem uma contribuição importante, especialmente em categorias como o DT. E o Tema, embora também desempenhe um papel relevante, tem uma contribuição mais isolada e menor, destacando a importância de considerar múltiplas variáveis no modelo de classificação.

A presença de correlações baixas ou moderadas entre as variáveis preditoras sugere que estas fornecem informações complementares, o que justifica a sua inclusão no modelo e potencializa a capacidade do modelo em realizar classificações precisas. Este resultado sublinha a importância de utilizar uma abordagem multivariada na análise de *corpora* textuais, onde diferentes dimensões — como o contexto de produção (Atividade), o conteúdo (Tema) e os tipos discursivos — são consideradas de forma integrada para maximizar a precisão do modelo.

A tabela de previsões (Anexo 8) fornece a previsão e a explicação para essa previsão, para cada uma das linhas[53] do conjunto de reserva (*hold-out*). No contexto dos modelos de classificação, as previsões referem-se às categorias ou classes atribuídas pelo modelo a novas observações, com base nos padrões aprendidos durante o processo de treino. Quando um modelo de classificação recebe uma nova observação (ou exemplo) com um conjunto de atributos ou variáveis de entrada, ele calcula a probabilidade de essa observação pertencer a cada uma das classes possíveis. A classe com a maior probabilidade é então atribuída como a previsão final. A tabela de previsões gerada pelo *RapidMiner* contém as seguintes colunas:

- **Atividade, Género, Tema, TD,**: Variáveis descritivas que contextualizam as instâncias.
- ***confidence* (Notícia), *confidence* (Comentário)**: Confiança nas previsões para cada classe ("Notícia" ou "Comentário").
- ***prediction* (Género)**: Classe prevista pelo modelo.
- ***cost***[54]: Custo associado à previsão.

De acordo com os dados obtidos, a maioria das instâncias está associada ao género "Comentário" como verdadeiro, mas o modelo prevê "Notícia" em vários casos com alta confiança. As previsões para "Notícia" apresentam confiança alta (≥ 0,7) na maioria das vezes, e para "Comentário", a confiança é frequentemente de 0,8, o que demonstra uma forte tendência do modelo para identificar corretamente esta classe. Relativamente ao custo associado, é maior (até 0,7) para instâncias mal classificadas, assim como nas previsões onde o modelo tem baixa confiança e classifica

[53] A tabela de previsões é constituída por 149 exemplos que correspondem à percentagem do *hold-out*: $373 \times 0{,}4 = 149{,}2$ $(aproximadamente\ 149\ linhas\ no\ hold-out)$

[54] O custo representa a consequência ou penalidade atribuída a uma previsão do modelo. Pode variar com o modo como o erro é definido ou as prioridades do problema. Embora, o modo de cálculo possa variar, o mais comum (e usado para este modelo) é:

$$Custo = 1 - Confiança\ na\ classe\ correta$$

incorretamente a instância. As instâncias com custo 0,0 representam previsões com 100% de acerto ou sem penalidade associada.

De um modo geral, podemos observar alguns exemplos de padrões que emergem da tabela:

- Em “Sociedade” com TD “N”, “Notícia” é prevista frequentemente com alta confiança (≥ 0,8).
- Outros temas, como “Política” e “Desporto”, também mostram alta confiança para “Notícia”.
- Casos como “Economia, Jornalística, DI, Comentário” têm confiança maior para “Comentário” (0,8), mas em algumas instâncias apresentam um custo elevado (≥ 0,6), o que mostra que o modelo errou com confiança alta.
- “Jornalística” é a atividade predominante em quase todas as instâncias, o que sugere que pode haver uma sobrerepresentação dessa atividade.
- “Tema” parece ter um papel mais secundário, com a previsão influenciada principalmente pela variável “TD”.

Esta tabela de previsões é importante porque nos mostra as vantagens e desvantagens do próprio modelo: se, por um lado, mostra que o GLM tem confiança alta para previsões corretas em ambas as classes e identifica bem padrões para temas como “Política” e “Sociedade” ao prever “Notícia”, por outro revela um custo alto em algumas previsões para “Comentário”, como nos casos “Economia” e “Literatura”, indicando que o modelo pode não captar bem a relação entre estes temas. O facto de algumas previsões terem confiança baixa ou moderada reflete incerteza nas previsões.

Relativamente ao custo associado aos erros de previsões, os custos altos (0.6 - 0.7) aparecem em diversas linhas da tabela, enquanto os custos baixos (≤ 0.2) também aparecem, mas com menos frequência. A presença de muitos custos elevados indica que o modelo não está totalmente robusto, e que tem dificuldade em identificar algumas classes, sobretudo aquelas que têm valores de confiança mais altos.

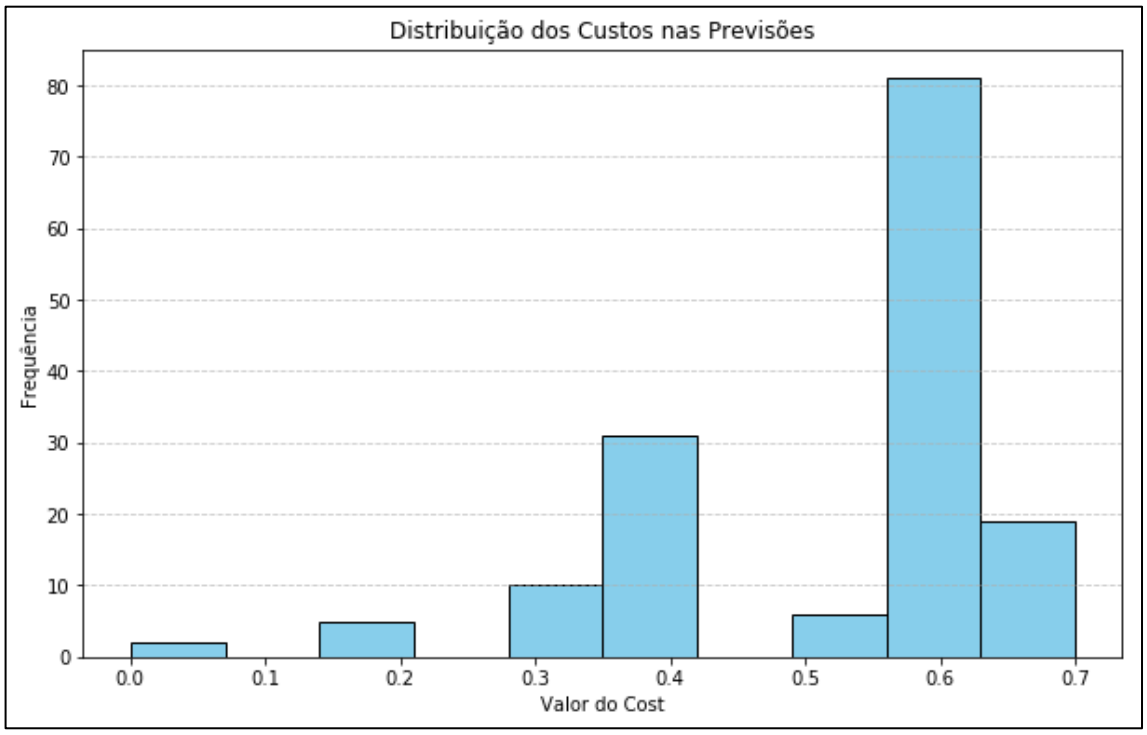


Figura 21: Distribuição dos custos associados às previsões.

De acordo com o gráfico da Figura 23, os custos associados às previsões concentram-se nos valores 0.6 (com aproximadamente 80 ocorrências) e 0.7 (com aproximadamente 20 ocorrências). Significa que estes valores mais elevados manifestam um de dois problemas: ou o modelo faz previsões erradas com um elevado grau de confiança, ou existe incerteza na classificação.

A tabela de previsões permite, portanto, identificar erros associados ao modelo e orientar a sua correção com vista à sua otimização:

- Interpretar os casos de maior custo para identificar padrões que o modelo pode não estar a captar;
- Verificar a representatividade dos temas “Economia” e “Literatura” no conjunto de treino;
- As previsões mostram um possível viés para “Comentário” devido à sua alta frequência nos dados, que pode ser corrigido com técnicas de calibragem;
- Executar ajustes para penalizar mais os erros com custo elevado, especialmente em previsões para “Comentário”.

Os resultados obtidos nesta secção evidenciam o bom desempenho do modelo de classificação baseado no Generalized Linear Model (GLM), com métricas que indicam elevada precisão na distinção entre os géneros "Notícia" e "Comentário". A precisão de 91,5% demonstra a robustez do modelo e a sua capacidade de separação entre as classes. A análise das métricas de *precision* (94,4%), *recall* (88,9%) e *specificity* (94,4%) confirma que o modelo apresenta um bom equilíbrio entre a identificação correta de instâncias positivas e a minimização de falsos positivos. Este equilíbrio é essencial no contexto da classificação de géneros textuais, assegurando uma correta distinção entre os diferentes tipos de textos analisados.

A análise dos pesos atribuídos pelo GLM às variáveis preditoras confirma a relevância da Atividade, que se destaca como o principal fator na classificação do género textual, com um peso de 0,313, superior ao dos TD (0,205) e do Tema (0,163). Estes resultados alinham-se com a perspetiva teórica que realça a abordagem descendente à análise dos textos. Embora o TD tenha um peso menor do que a Atividade, a sua contribuição ainda é significativa. Por outro lado, o Tema, com o menor peso entre as variáveis, revela uma influência mais complementar, sendo relevante em combinação com as outras variáveis preditoras.

A tabela de previsões gerada pelo modelo destaca a confiança elevada nas previsões corretas, especialmente para a classe "Notícia", com valores superiores a 0,7 na maioria dos casos. No entanto, observa-se a presença de custos elevados (0,6 - 0,7) em algumas previsões erradas, e os problemas identificados poderão ser corrigidos futuramente através de técnicas de calibragem para equilibrar a representação das diferentes atividades e temas no conjunto de treino.

De modo geral, os resultados reforçam a pertinência da abordagem metodológica adotada, evidenciando a importância de considerar múltiplas variáveis preditoras na análise de géneros textuais. A preponderância da Atividade confirma o seu papel fundamental para o género, enquanto a inclusão dos Tipos Discursivos e do Tema permite um modelo mais equilibrado e preciso. Esta análise do modelo de classificação valida a abordagem multivariada utilizada.

7. Reflexões finais

Este trabalho teve como ponto de partida duas questões centrais, as quais orientaram esta investigação:

(i) Será a prática textual do comentário um género relativamente estabilizado, com características próprias e fronteiras delimitadas relativamente a outros géneros textuais, ou estamos, pelo contrário, perante uma nebulosa de textos sem fronteiras nítidas?

(ii) Se, de facto, estamos perante um género textual estável, que características podemos elencar e como podemos validar essas características?

Em relação à primeira questão, a análise conduzida não conseguiu identificar de forma definitiva um género claramente estabilizado, mas sim uma diversidade de textos que coexistem sob a etiqueta de "comentário". Embora existam marcas linguísticas que indicam uma tendência para a homogeneização do género, as fronteiras entre o comentário e outras práticas discursivas, como a notícia, continuam imprecisas. Isto sugere que a prática textual do comentário não é um género estabilizado, mas sim uma prática comunicativa fluída e em constante adaptação aos contextos de produção e receção em que ocorre. Portanto, a nossa hipótese inicial de que o comentário poderia ser considerado uma nebulosa de textos sem fronteiras nítidas parece ser corroborada pelos dados analisados.

No que respeita à segunda questão, a caracterização das propriedades linguísticas e discursivas do comentário foi realizada com base em dados quantitativos e qualitativos, utilizando ferramentas de *text mining* e *data mining*. A análise revelou que o comentário apresenta certas características distintivas, como a predominância de pronomes de primeira e segunda pessoa e advérbios de lugar e tempo com valor deítico, que indicam uma forte relação com o ponto de vista do enunciador. Além disso, a análise mostrou que o comentário é frequentemente produzido em contextos de comunicação mediática e social, o que reforça a sua natureza dinâmica e adaptativa às necessidades do público-alvo. No entanto, a validação dessas características como traços exclusivos do comentário não foi totalmente conclusiva, o que aponta para a necessidade de mais estudos para fundamentar estas marcas como características definidoras do género.

A metodologia adotada, baseada na combinação de análise linguística e ferramentas de *data mining*, permitiu explorar as interações entre as unidades linguísticas e os tipos discursivos de forma sistemática. Contudo, a análise quantitativa não foi suficiente para identificar padrões claros entre os diferentes *subcorpora*, pois os testes estatísticos de relevância não mostraram diferenças significativas entre as médias de ocorrência das unidades linguísticas. Entretanto, quando as variáveis foram analisadas em conjunto, como evidenciado pelos dendrogramas obtidos, emergiram padrões de agrupamento que sugerem a existência de traços distintivos entre o comentário e outros géneros textuais, como a notícia. Este resultado sublinha a importância de uma análise qualitativa e holística, que ultrapasse as médias estatísticas e considere a coocorrência das variáveis, permitindo identificar padrões mais complexos e significativos.

Os dados obtidos através do modelo de classificação ilustram como as ferramentas de *text mining* e *data mining* podem abrir novas direções para a pesquisa de géneros textuais. A análise quantitativa, combinada com a interpretação qualitativa dos dados, mostra que é possível extrair informações relevantes sobre a estrutura e a organização dos textos de comentário, e como essas ferramentas podem ser aplicadas para identificar características linguísticas que não seriam facilmente observáveis por uma análise puramente manual. Contudo, os resultados também indicam que o modelo de classificação utilizado é apenas uma aproximação inicial, e que há margem para aperfeiçoamento, nomeadamente na integração de outros marcadores de género que poderiam enriquecer a análise e a identificação do comentário como género textual.

Por fim, a nossa investigação não apenas confirma a complexidade da identificação e classificação dos géneros textuais, mas também aponta para a necessidade de abordagens mais flexíveis e dinâmicas no uso de ferramentas automatizadas de análise de texto. A integração do quadro teórico do ISD (Interacionismo Sociodiscursivo) no *data mining* revela-se útil e pertinente para a análise dos textos, permitindo explorar as interações entre os elementos linguísticos e as práticas discursivas. A metodologia proposta abre, assim, novas possibilidades para o estudo de géneros textuais, sobretudo aqueles que se caracterizam por estruturas híbridas e complexas, e sugere que futuras investigações podem aprofundar a aplicação

destas ferramentas, ampliando o alcance da análise de texto para contextos e géneros ainda mais diversificados.

Deste modo, este trabalho também contribui significativamente para a exploração de novas formas de visualizar dados na análise do texto, utilizando ferramentas de *text mining* e *data mining* que integrem os níveis meso e macro da análise textual. A matriz de correlação permitiu uma visualização clara das relações entre os tipos discursivos (TD) no *corpus*, revelando como a homogeneidade ou heterogeneidade da distribuição dos TD influencia as correlações. Ao focar nas coocorrências entre os TD, a análise revelou padrões complexos que seriam difíceis de identificar através de métodos manuais, oferecendo uma nova perspetiva sobre a organização interna dos textos. A utilização de dendrogramas para análise de *clustering* proporcionou uma visualização eficaz da forma como os textos se agrupam, permitindo que os dados fossem observados de forma estruturada. Esta abordagem não só revelou como os fenómenos microlinguísticos característicos de cada texto podem ser agrupados, mas também mostrou como os padrões podem ser mapeados de maneira visual, proporcionando uma nova forma de ver a distribuição unidades linguísticas. Assim, ao aplicar técnicas de *data mining* na análise textual, este trabalho demonstra como a visualização de dados pode revelar novas camadas de significado, descobrir novos padrões e oferecer uma análise mais detalhada da dinâmica dos géneros textuais, facilitando a interpretação e a compreensão dos dados em contextos diversificados.

Em conclusão, este estudo contribui para o avanço da análise de géneros textuais, ao integrar metodologias inovadoras de análise quantitativa com a teoria do ISD, e aponta direções importantes para a otimização dessas abordagens no futuro.

Referências Bibliográficas

Aaronson, S. (2001). Stylometric clustering: a comparison of data-driven and syntactic features. *Manuscript. Http://Www. Cs. Berkeley. Edu/Aaronson/Sc. Doc*, 20 p.

Adam, J.-M. (1997). Unités rédactionnelles et genres discursifs: cadre général pour une approche de la presse écrite. *Pratiques*, 3–18. Retrieved from http://www.pratiques-cresef.com/p094_ad1.pdf

Adam, J.-M. (2001). Types de textes ou genres de discours ? Comment classer les textes qui disent de et comment faire ? *Langages*, (141), 10–27. https://doi.org/10.3406/lgge.2001.872

Adam, J.-M. (2008). *A Linguística textual. Introdução à análise dos discursos*. São Paulo: Cortez.

Adam, J. (1992). *Les textes: types et prototypes*. [Paris]: Éditions Nathan.

Adam, J. (2001). En finir avec les types de texte. In M. Ballabriga (Ed.), *Analyse des discours : types et genres : communication et interprétation* (pp. 25–43). Toulouse: Editions universitaires du Sud.

Adam, J. (2012). *Analyse textuelle des discours: niveaux ou plans d´analyse* (Vol. 14).

Adam, J. (2013). Problèmes du texte. *Pré Publications*, (200).

Biber, D. (1989). A typology of English texts. *Linguistics, 27*(1), 3–43.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Biber, D. (1995a). *Dimensions of register variation*. Cambridge: Cambridge University Press.

Biber, D. (1995b). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (2004). Conversation text types : A multi-dimensional analysis. *Le Poids Des Mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*, *1*, 15–34. Retrieved from http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_000.pdf

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly.

Blache, P., Ferré, G., Rauzy, S., Blache, P., Ferré, G., Rauzy, S., … Scheme, C. (2008, July). An XML Coding Scheme for Multimodal Corpus Annotation. *Corpus Linguistics*, 1–17.

Bronckart, J.-P. (2008). Discussion de quelques concepts pour une approche praxéologique du langage. In J. Durand, B. Habert, & B. Lacks (Eds.), *Congrès Mondial de Linguistique Française - CMLF'08* (pp. 861–867). Paris: Institut de Linguistique Française. https://doi.org/10.1051/cmlf08313

Bronckart, Jean-Paul. (1988). *Le fonctionnement des discours. Un modèle psychologique et une méthode d'analyse*. *Pratiques* (Vol. 58). Neuchâtel: Delachaux et Niestlé.

Bronckart, Jean-Paul. (1997). *Activité langagière, textes et discours. Pour un interactionisme socio-discursif*. Paris: Delachaux et Niestlé.

Bronckart, Jean-Paul. (2004). Les genres de textes et leur contribution au développement psychologique. *Langages*, *n° 153*(1), 98–108. https://doi.org/10.3917/lang.153.0098

Bronckart, Jean-Paul. (2008). Genres de textes, types de discours et "degrés" de langue : hommage à François Rastier. *Texto !*, *XIII*(1/2). Retrieved from http://www.revue-texto.net/index.php?id=86

Bronckart, Jean-Paul. (2010). La vie des signes en questions: des textes aux langues, et retour. In A. M. et alii Brito (Ed.), *Textos Seleccionados, XXV Encontro Nacional da Associação Portuguesa de Linguística* (pp. 11–41). Porto: APL.

Broucker, J. de. (1995). *Pratique de l'information et écritures journalistiques*. Paris: CFPJ.

Calabrese, L. (2019). Le commentaire : continuités et mutations d'un outil au service de la lecture et de l'écriture. In L. Calabrese (Ed.), *Revue de linguistique française et d'analyse du discours* (pp. 7–28). Paris: Editions L'Harmattan.

Calabrese, L., & Jenard, J. (2018). Talking about News. A Comparison of readers' comments on Facebook and news websites. *French Journal For Media Research*, (10).

Chein, F. (2019). *Introdução aos modelos de regressão linear Metodologias*. Brasília: Enap.

Correia, C. N., & Pereira, S. (2014). Sobre a construção do espaço e do tempo: as formas cá e lá em português europeu. In A. Fiéis, M. Lobo, & A. Madeira (Eds.), *O universal e o particular: uma vida a comparar* (pp. 103–115). Lisboa: Colibri.

Costa, F., & Branco, A. (2012). *Aspectual Type and Temporal Relation Classification*.

Coutinho, M. A. (1999). *Texto(s) e competência textual*. Faculdade de Ciências Sociais e Humanas - UNL.

Coutinho, M. A. (2006). O texto como objecto empírico: consequências e desafios para a lingüística. *Veredas (Revista de Estudos Linguísticos - UFJF)*, *10*(1–2), 1–13. Retrieved from http://www.ufjf.br/revistaveredas/files/2009/12/artigo076.pdf

Coutinho, M. A. (2019). *Texto e[m] Linguística*. Lisboa: Colibri.

Coutinho, M. A., & Miranda, F. (2009). To describe genres: Problems and strategies. *Genre in a Changing World*, 35–55. Retrieved from http://wac.colostate.edu/books/genre/ http://www.parlorpress.com/genre

Cristea, D., & Butnariu, C. (2009). Annotating Discontinuous Structures in XML: the Multiword Case.

Crowston, K., & Kwasnik, B. H. (2004). A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality", Proceedings of the 37th Hawaii International Conference on System Science (HICSS '04). In *Proceedings of the 37th Hawaii International Conference on System Science (HICSS '04).*

De Beaugrande, R., & Dressler, W. U. (1981). Introduction to Text Linguistics. *Introduction to Text Linguistics*. https://doi.org/10.4324/9781315835839

Dickey, E. (2007). *Ancient Greek Scholarship. A guide to Finding, Reading, and Understanding Scholia, Commentaries, Lexica, and Grammatical Treatises, from Their Beginnings to the Byzantine Period*. Oxford/New York: Oxford University Press.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook*. *Imagine*. Cambridge: Cambridge University Press. Retrieved from http://arxiv.org/abs/1011.1669%5Cnhttp://dx.doi.org/10.1088/1751-8113/44/8/085201

Field, A. (2017). *Discovering Statistics Using IBM SPSS Statistics*. London: Sage Publications.

Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*. https://doi.org/10.1002/asi.20427

Fonseca, F. I. (1992). *Deixis, tempo e narração*. Porto: Fundação Eng. António de Almeida.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Goecke, D., Liingen, H., Metzing, D., & Stiihrenherg, M. (2010). Different Views on Mark up Distinguishing Levels and Layers. In A. Witt & D. Metzing (Eds.), *Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology* (pp. 1–22). Dordrecht: Springer.

Gonçalves, M., & Carrilho, J. (2020). Comentando comentários: questões de texto, género e corpus. *Revista Da Associação Portuguesa de Linguística*, (7), 191–208. https://doi.org/10.26334/2183-9077/rapln7ano2020a12

Gonçalves, M., & Magalhães, M. (2019). Corpus e géneros textuais nas práticas de divulgação de ciência ou as novas hierarquias na construção do conhecimento. *Revista Da Associação Portuguesa de Linguística*, (5), 145–157. https://doi.org/10.26334/2183-9077/rapln5ano2019a11

Gonçalves, M., & Miranda, F. (2008). Analyse Textuelle, Analyse De Genres : quelles relations, quels instruments? In *Autour des langues et du langage*. Presses Universitaires de Grenoble.

Goody, J. (1977). *The Domestication of the Savage Mind*. Oxford University Press.

Habermas, J. (1987). *Théorie de l'agir communicationnel* (Vol. 1). Paris: Fayard.

Hagège, C., Baptista, J., & Mamede, N. (2010). Caracterização e Processamento de Expressões Temporais em Português. *Linguamatica*, *2*.

Hardie, A. (2014). Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal*, *38*(1), 73–103. https://doi.org/10.2478/icame-2014-0004

Ihlstrom, C., & Akesson, M. (2004). Genre characteristics - a front page analysis of 85 Swedish online newspapers. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (p. 10 pp.). IEEE. https://doi.org/10.1109/HICSS.2004.1265274

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2022). An introduction to statistical

learning with applications in R. *Statistical Theory and Related Fields*, *6*(1), 87–87. https://doi.org/10.1080/24754269.2021.1980261

Jorge, N. (2014). *O género memórias. Análise linguística e perspetiva didática*. Faculdade de ciências sociais e humanas - UNL.

Kessler, B., Nunberg, G., & Schutze, H. (1997). Automatic Detection of Text Genre. In *ACL '98 Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 32–38). https://doi.org/10.3115/976909.979622

Kilgarriff, A., Tugwell, D., Rychly, P., & Smrz, P. (2004). The sketch engine. In *Proceedings of Euralex*. Lorient.

Kuhn, T. Z. (2019). *A design proposal of an online corpus-driven dictionary of Portuguese for university students*. *Journal of Portuguese Linguistics*. https://doi.org/10.5334/jpl.209

Laidouni, N. (2019). *LITTÉRATURE ET PRESSE : UNE ÉTUDE DE COMPRÉHENSION DES TEXTES (EN CONTEXTE LIBANAIS)*. Université Côte d'Azur.

Lee, D. Y. W. (2001). Genres, Registers, Text types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology*, *5*(3), 37–72. https://doi.org/10.1016/S1364-6613(00)01594-1

Lohfink, G. (1973). Kommentar als Gattung, *72*(1947), 203–208.

Lyons, J. (1977). Deixis, Space and Time - Cap.15. In *Semantics* (pp. 636–724). Cambridge: CUP.

Magalhães, M., & Gonçalves, M. (2021). A *deixis* : uma proposta de anotação em XML no âmbito do texto, 103–116.

Maingueneau, D. (2007). Genres de discours et modes de généricité. *Le Français Aujourd'hui*, *159*(4), 29. https://doi.org/10.3917/lfa.159.0029

Malrieu, D., & Rastier, F. (2001). Genres et variations morphosyntaxiques. *Revue TAL : Traitement Automatique Des Langues*, *42*(2), 547–577.

Manguel, A. (2020). *Uma História da Leitura*. Lisboa: Tinta da China.

Marx, K. (2015). *O Capital* (Vol. 1). Boitempo.

Mateus, M. H. M., Brito, A. M., Duarte, I., & Faria, I. H. (2006). *Gramática da Língua Portuguesa* (7ª). Lisboa: Caminho.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics: an introduction* (2.º). Edinburgh: Edinburgh University Press.

McEnery, Tom, & Wilson, A. (2001). *Corpus Linguistics: an introduction* (2nd ed.). Edinburh: Edinburgh University Press.

Miranda, F. (2007). Marcadores de gênero: uma pista para identificar a ficcionalizaçãode gêneros textuais. In *Proceedings of the 4th SIGET – International Symposium on Genre Studies* (pp. 1045–1055).

Miranda, F. (2008). Géneros De Texto E Tipos De Discurso Na Perspectiva Do Interaccionismo Sociodiscursivo: Que Relações? *Estudos Linguísticos*, 81–100.

Miranda, F. (2010). *Textos e géneros em diálogo. Uma abordagem linguística da intertextualização*. Lisboa: FCG/FCT.

Piaget, J. (1950). *Introduction a l'épistémologie génétique*. Paris: PUF.

Pereira, M. C. (2009). Breve Abordagem Semântica e Pragmática de Aqui , Aí e Ali. *ELingUp [Centro*, *1*(1), 60–80.

Perkins, J. (2014). *Python 3 Text Processing With NLTK 3 Cookbook*. *Python 3 Text Processing With NLTK 3 Cookbook*. https://doi.org/10.1017/CBO9781107415324.004

Rastier, F. (2001). *Arts et sciences du texte*. Paris: PUF.

Rastier, F. (2011). *La mesure et le grain. Sémantique de corpus*. Paris: Honoré Champion.

Rehm, G. (2002). Towards automatic Web genre identification: a corpus-based approach in the domain of academia by example of the Academic's Personal Homepage. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (pp. 1143–1152). IEEE Comput. Soc. https://doi.org/10.1109/HICSS.2002.994036

Rico, C. (2003). Aux sources de l'herméneutique occidentale :les premiers commentaires dans les traditions grecque, juive et chrétienne. *Babel*, (7), 7–52. https://doi.org/10.4000/babel.1404

Rocha, P., & Santos, D. (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *V Encontro Para o Processamento Computacional Da Língua Portuguesa Escrita e Falada (PROPOR 2000)*, 131–140.

Santini, M. (2004). State-of-the-Art on Automatic Genre Identification. *Technology*.

Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=16002295292928789832related:SA3t_ciSE94J

Santini, M. (2005). Linguistic facets for genre and text type identification: A description of linguistically-motivated features. *ITRI Report Series: ITRI-05*, 1–41. Retrieved from http://www.nltg.brighton.ac.uk/home/Marina.Santini/linguistic_facets_tech_rep.pdf

Santini, M. (2006). Web Pages, Text Types, and Linguistic Features: Some Issues. *ICAME Journal*, *30*. https://doi.org/10.18500/2311-0740-2019-1-21-22-33

Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, *8*, 65–70. https://doi.org/10.1.1.132.6973

Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to data mining*. Pearson.

TEI Consortium (Ed.). (2002). A Gentle Introduction to XML. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange (Text Encoding Initiative Consortium* (pp. 1–26).

Teixeira, J. (2005). De Cá para Lá e de aqui para aí: Rede De Valores Semânticos Dos Marcadores Espaciais cá/lá/(acolá) e aqui/aí/ali. In G. M. Rio-Torto, O. M. Figueiredo, & F. Silva (Eds.), *Estudos em Homenagem ao Professor Doutor Mário Vilela* (Vol. 1, pp. 449–460). Porto: Faculdade de Letras da Universidade do Porto.

Valentim, H. T. (2015). Deixis in European Portuguese: Representation and Reference Construction. In K. J. F. Da Milano (Ed.), *Manual of Deixis in Romance Languages* (pp. 247–314). Berlin/Bos: Mouton De Gruyter.

Valentim, H. T., & Gonçalves, M. (2021). intensificação em português europeu - Algumas configurações linguísticas em comentários em linha. *Estudios Románicos*, *30*, 103–120. https://doi.org/10.6018/ER.471941

Volochinov, V. N. (1977). *Volochinov, V.N.Le marxisme et la philosophie du langage. Paris : Minuit.* Paris: Minuit.

Vygotsky, L. S. (1997). *Pensée et langage*. Paris: La Dispute.

Werlich, E. (1976). *A Text Grammar of English, , Heidelberg (Germany).* Heidelberg: Quelle & Meyer.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition* (2nd ed.). Morgan Kaufman.

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Anexo 1: textos exemplares

TEXTO 1

Onde ver arquitectura?

Comentário
Isabel Salema

Se a qualidade e a projecção da arquitectura nacional já valeram dois Prémios Pritzker a Portugal, a Siza e a Souto de Moura, a abertura de um museu dedicado à disciplina tem sido um assunto constantemente adiado. Um dos argumentos mais usados na discussão, que já tem seguramente mais de uma década, é que não se trata de um tema fácil de expor. Os menos entusiastas logo acrescentam que o melhor sítio para ver arquitectura é mesmo a própria cidade onde ela se ergue todos os dias.

Entre os museus públicos, o Centro Cultural de Belém destaca-se em Lisboa com um programa regular de exposições dedicado à arquitectura, enquanto Serralves, no Porto, tem pontualmente organizado mostras dedicadas ao tema. Se o novo MAAT inclui arquitectura no seu nome, a sua área de actuação é muito mais abrangente e muito ancorada nas artes visuais. Sem ser exaustiva, a Gulbenkian também tem tido alguma actividade expositiva ligada à arquitectura, tal como a Casa da Cerca, em Almada, mas uma maior visibilidade, e uma maior ambição, chegam só de três em três anos com a programação da Trienal de Arquitectura.

Nuno Sampaio, que reivindica com a Casa de Arquitectura ter conseguido fazer em Matosinhos o primeiro museu só de arquitectura, lembra que esta é a única área relevante da Cultura em que o Estado não está presente com uma instituição de programação que lhe seja inteiramente dedicada, quando internacionalmente é reconhecida a sua grande qualidade.

Agora que a Casa de Arquitectura abrirá a Norte no próximo Verão, com uma ambição internacional, como mostra esta apresentação em Veneza, é preciso que consiga ultrapassar a divisão Norte-Sul, às vezes demasiado presente na arquitectura portuguesa. E que não caia em armadilhas regionalistas, privilegiando demasiado a arquitectura da chamada Escola do Porto. Mas, mais do que da Casa, isso depende também dos arquitectos portugueses e da sua vontade de trabalhar em conjunto num projecto nacional.

TEXTO 2

Comentário do especialista

CES Las Vegas

Chegámos à CES 2015, Consumer Electronics Show, de Las Vegas, com alguma expectativa sobre os televisores *quantum dot* LED. O entusiasmo esmoreceu ao confirmarmos que o dito novo tipo de iluminação de ecrã já tinha sido usado nalguns modelos da série Triluminos da Sony, em 2013. Tínhamos testado alguns e confrontámos a teoria com os resultados obtidos na altura, em laboratório. Cores mais nítidas, naturais e com uma maior gama: a promessa dos microscópicos *quantum dots*, com 2 a 10 nanómetros, baseia-se na capacidade de reproduzir as cores com maior precisão. As suas minúsculas dimensões permitem comprimentos de onda menores, traduzindo-se em cores mais precisas. Se expostos à luz azul emitida pelos LED, os *quantum dots* convertem alguma dessa luz, na origem, para verde e vermelho. O resultado na imagem abrange as três cores primárias.

António Alves
PRODUTOS
E SERVIÇOS

"Os poucos modelos *quantum dot* LED analisados dividiram um pouco as opiniões no painel, o que não sucedeu nos LCD LED"

Quantum dot LED divide opiniões

Escrutinámos os resultados dos modelos da Sony com *quantum dot* LED. Ao contrário dos televisores LCD LED, nos *quantum dot*, a perceção das cores divide um pouco o painel que visualiza as imagens dos televisores no nosso teste. Alguns dão boa nota às cores, mas outros relatam desvios nos tons. Sentados frente a um ecrã *quantum dot* para ver uma imagem só com as cores mais puras, ficaríamos impressionados com a tecnologia. Mas nas imagens de televisão, como na vida, não vemos só cores puras. Também há misturas de cores ou cores pastel, como nos tons de pele. É aí que a diferença de opiniões começa.

TEXTO 3

Comentário do especialista

Drones encolhem

Na feira eletrónica de 2015, em Las Vegas, os drones reclamaram o bilhete de entrada desses equipamentos ao apostar nas *selfies* ou autorretratos. Fotógrafos e profissionais de vídeo já "atrelavam" desde *action cams* (câmaras de vídeo aventura), máquinas fotográficas reflex a drones para obter ângulos incríveis, em aparelhos mais caros e volumosos. Mas os consumidores também estão na mira dos fabricantes para usos menos ambiciosos. Os fabricantes destas pequenas aeronaves integram câmaras que alegam uma qualidade melhor e estabilizador de imagem em drones. Mais: alguns apostam nas *selfies*.

António Alves
PRODUTOS
E SERVIÇOS

"A tecnologia *follow me* ("segue-me") está a dar cartas na área dos drones: permite a um drone acompanhar e filmar ou fotografar de forma automática o utilizador"

Vamos tirar uma selfie

A tecnologia *follow me* ("segue-me") permite a um drone acompanhar e filmar ou fotografar de forma automática o utilizador. É ideal para vídeos e fotografias aéreas de desportos radicais. Há um investimento aparente na qualidade de imagem. O drone Zano, por exemplo, pode ficar numa posição fixa no ar enquanto captura a imagem ou ser programado para seguir o movimento da sua aventura. Compacto e leve, rivaliza com os suportes extensíveis para capturar fotografias com telemóveis, câmaras e *action cams*. Tão pequeno que pode ser usado no pulso, o ainda projeto "Nixie" dá o salto para um drone que é um acessório: basta atirá-lo e a força determina a distância a que se coloca do utilizador; através do reconhecimento facial, vira-se para o "dono" e captura o autorretrato. Da DJI, a marca do Phantom, o drone Inspire 1 apresenta uma câmara integrada, que grava em resolução 4K, e pode ser controlado por um ou dois *smartphones* ao mesmo tempo. Um utilizador controla o drone e o outro a posição da câmara.

Drone Zano com vários modos de voo

TEXTO 4

COMENTÁRIO POLÍTICO

Marques Mendes: "O PSD está muito reduzido a Passos Coelho e a Maria Luís Albuquerque"

14/8/2016, 22:11 779 41

O comentador acredita que Pedro Passos Coelho tem, primeiramente, de vencer as autárquicas, mudar de discurso e apostar noutras figuras. Ou então pode haver uma "indesejável crise interna".

Partilhe

"Se ganhar as eleições autárquicas fica mais bem preparado para ganhar as eleições legislativas"
André Antunes

Autor

Miguel Santos
Miguel_SantosC
Email

Mais sobre

COMENTÁRIO POLÍTICO
POLÍTICA

O PSD precisa de mudar a agulha. Pelo menos, é a esta a opinião de Luís Marques Mendes expressa no dia em que o PSD faz a *rentrée* política no Pontal. O ex-líder social-democrata considera que, "em termos públicos", o partido está muito reduzido a Passos Coelho e a Maria Luís Albuquerque". Os sociais-democratas, diz o comentador, tinham mais a ganhar em "criar uma equipa de porta-vozes" que falasse sobre "as várias áreas setoriais".

E este não é o único desafio que a atual direção do PSD enfrenta, considera Marques Mendes. No habitual espaço de comentário, na SIC, o ex-presidente social-democrata apontou o caminho: o PSD tem de "ganhar as eleições autárquicas" e apostar em "grandes candidatos" para os principais centros urbanos. "Se ganhar as eleições autárquicas fica mais bem preparado para ganhar as eleições legislativas". Ou então pode haver uma "indesejável crise interna e de liderança", foi avisando o comentador.

Mas só isto não chega. Luís Marques Mendes considera que é tempo de Pedro Passos Coelho "mudar o discurso", até ao momento "muito concentrado no passado", "derrotista" e "centrado nas questões financeiras". "As pessoas querem ouvir falar no futuro", um discurso "gerador de "esperança" e "mais abrangente", voltado para questões "sociais, culturais e científicas", repetiu o comentador.

Ainda assim, não foi apenas o PSD a merecer reparos do social-democrata: o comentador não poupou críticas à forma como o Governo socialista geriu o dossier "Caixa Geral de Depósitos". "Tiros nos pés", "leviandades" e "muita precipitação", apontou Marques Mendes. O comentador acredita que o facto de a Caixa ter ficado "oito meses sem administração" contribuiu decisivamente para a "degradação da imagem do banco público".

Além disso, insistiu, o facto de o Banco Central Europeu ter, de acordo com o jornal Público, exigido a redução do número de novos administradores de 19 — como pretendia o Executivo português — para 11 parece uma "brincadeira de crianças", constitui uma "verdadeira humilhação" para as pessoas convidadas e é uma "derrota para o Ministério das Finanças".

Relacionado

- O coração do PSD ainda bate por Passos?

PARTILHE

COMENTE

41 Comente e partilhe as suas ideias

SUGIRA

Proponha uma correção, sugira uma pista:
msantos@observador.pt

Anexo 2: grelhas de análise dos textos exemplares

1. Mecanismos de realização textual

(I) Aspetos situacionais

Parâmetros Físicos				
	Produtor	**Recetor**	**Local**	**Temporalidade**
Texto 1	Isabel Salema	n. identificado	n. identificado	26/11/2016
Texto 2	António Alves	n. identificado	n. identificado	02/2015
Texto 3	António Alves	n. identificado	n. identificado	03/2015
Texto 4	Miguel Santos	n. identificado	n. identificado	14/08/2016

Parâmetros sociossubjetivos				
	Enunciador	**Destinatário**	**Finalidade**	**Quadro Social de circulação**
Texto 1	Jornalista Especializado (cultura)	Público em geral / especializado	○ Apresentar um acontecimento E/OU produto. ○ Refletir sobre acontecimento E/OU produto	Jornal diário nacional
Texto 2	Jornalista Especializado (tecnologia)	Público em geral / especializado	○ Apresentar um acontecimento E/OU produto. ○ Refletir sobre acontecimento E/OU produto	Revista mensal nacional
Texto 3	Jornalista Especializado (tecnologia)	Público em geral / especializado	○ Apresentar um acontecimento E/OU produto. ○ Refletir sobre acontecimento E/OU produto	Revista mensal nacional
Texto 4	Jornalista Especializado (política)	Público em geral / especializado	○ Dar a conhecer a opinião de outrem	Sítio de informação nacional

(II) Dimensão semiolinguística

2. Mecanismos Temáticos

	Temas	**Léxico**	**Coesão temporal e aspetual:**
Texto 1	Comentário sobre a musealização da arquitetura em Portugal E anúncio da abertura da Casa de Arquitetura.	Abundância de Nomes Próprios: toponímicos (Portugal, Lisboa, Porto...), Entidades (Prémios Pritzker; Centro Cultural de Belém...), Onomásticos (Siza, Souto de Moura...).	Presença de formas no conjuntivo associadas a frases impessoais que expressam opinião/conselhos ("é preciso que consiga..."; "e que não caia...") Presente do Indicativo e Pretérito Perfeito Composto para descrever situação presente.
Texto 2	Apresentação da CES 2015 e dos novos televisores.	Vocabulário especializado da eletrónica de consumo (*dot, LED,* luz azul...)	Organização temporal em torno do evento passado com recurso predominante do Pretérito Perfeito Simples do Indicativo (chegámos, esmoreceu...) e do Pretérito Mais-Que-Perfeito do Indicativo (tinha sido, tínhamos testado...). Presente do Indicativo com valor gnómico ("não vemos só cores puras", "Também há misturas de cor"...)
Texto 3	Apresentação da CES 2015 e dos novos drones.	Vocabulário especializado da eletrónica de consumo (drones, resolução 4K...)	Uso predominante do Pretérito Perfeito Simples e do Imperfeito e do Presente do Indicativo com valor gnómico.
Texto 4	Comentário político de Luís Marques Mendes. Tema apresentado por etiqueta peritextual (comentário político)	Vocabulário predominantemente associado à atividade política	Predomínio quase exclusivo do Presente do Indicativo com valor deítico.

3. Mecanismos Enunciativos

	Referência Temporal	Referência Pessoal	Referência espacial	Responsabilização
Texto 1	Ancoragem situacional (data de publicação no peritexto do jornal "26 de novembro de 2016"); marcadores adverbiais ("Agora", "No próximo verão") .	Ausência de deíticos pessoais.	Ausência de deíticos espaciais.	Indicação do autor no peritexto.
Texto 2	Ancoragem situacional (data de publicação no peritexto do jornal "fevereiro de 2015").	Deíticos pessoais (1ª pessoa do plural) através da morfologia verbal (Chegámos, escrutinámos...)	Ancoragem situacional (localizadores autónomos "em Las Vegas")	Indicação do autor e estatuto social no peritexto ("Produtos e serviços"; "comentário do especialista")
Texto 3	Ancoragem situacional (data de publicação no peritexto do jornal "março de 2015").	Ausência de deíticos pessoais	Ancoragem situacional (localizadores autónomos "em Las Vegas")	Indicação do autor e estatuto social no peritexto ("Produtos e serviços""; "comentário do especialista")
Texto 4	Ancoragem situacional (data de publicação no peritexto do jornal "14 de agosto de 2016"). Localizadores autónomos ("No dia em que o PSD faz a rentrée política no Pontal")	Ausência de deíticos pessoais	Ancoragem situacional (localizadores autónomos "na SIC")	Indicação do autor no peritexto.

4. Marcadores Composicionais

	Plano de texto	Tipos discursivos

Texto 1	• Unidades verbais (três colunas de texto), não verbais (fotografia da autora). • Título (Onde ver arquitectura?), etiqueta de género (comentário) e nome do autor (Isabel Salema).	*Discurso Teórico (dominante):* • Frases declarativas; • Predomínio do tempo Presente do Indicativo. • Predomínio da 3ª pessoa. • Conjunção das coordenadas do discurso em relação ao mundo real. • Autonomia em relação ato de produção. *Discurso Interativo (encaixe) no último parágrafo:* • Predomínio do modo Conjuntivo, construções com valor de futuro ("é preciso que consiga"; "E que não caia") • Deíticos temporais • Presença de auxiliares de modo (ser preciso) • Conjunção das coordenadas do discurso em relação ao mundo real. • Implicação em relação ato de produção
Texto 2	• Unidades verbais (uma coluna de texto), não verbais (fotografia do autor, do autor no evento). • Etiqueta de secção (comentário do especialista); Título (CES Las Vegas). Nome do autor e responsabilidade (António Alves; Produtos e Serviços) e citação do texto.	*Discurso Teórico:* • Frases declarativas; • Predomínio do tempo Presente do Indicativo. • Predomínio da 3ª pessoa. • Conjunção das coordenadas do discurso em relação ao mundo real. • Autonomia em relação ato de produção. *Discurso Interativo (intercalado):* • Predomínio do modo Indicativo (Presente e Pretérito Perfeito Simples) • Uso da 1ª pessoa do plural (chegámos, confrontámos, escrutinámos) • Conjunção das coordenadas do discurso em relação ao mundo real. • Implicação em relação ato de produção
Texto 3	• Unidades verbais (uma coluna de texto), não verbais (fotografia do autor, do produto em análise). • Etiqueta de secção (comentário do especialista); Título (CES Las Vegas). Nome do autor e responsabilidade (António Alves; Produtos e Serviços) e citação do texto.	*Discurso Teórico:* • Frases declarativas; • Predomínio do modo Indicativo (Presente e Pretérito Perfeito Simples) • Predomínio da 3ª pessoa (singular e plural). • Conjunção das coordenadas do discurso em relação ao mundo real. • Autonomia em relação ato de produção.

Texto 4	• Unidades verbais (uma coluna de texto), não verbais (fotografia do comentador). • Etiqueta de secção (comentário político); Título (Marques Mendes: "O PSD está muito reduzido a Passos Coelho e a Maria Luís Albuquerque). Nome do autor (Miguel Santos).	*Discurso Teórico (dominante):* • Frases declarativas; • Predomínio do modo Indicativo (Presente e Pretérito Perfeito Simples) • Predomínio da 3ª pessoa (singular e plural) e de construções impessoais. • Presença de modalizações lógicas e de auxiliares (querer, poder, ter de). • Conjunção das coordenadas do discurso em relação ao mundo real. • Autonomia em relação ato de produção. *Discurso Interativo (encaixe) intercalado:* • Predomínio do Presente Indicativo com valor deítico • Presença de auxiliares de modo (precisar, ter de). • Presença de pontuação (uso de aspas) que remetem para a interação verbal (real ou encenada). • Conjunção das coordenadas do discurso em relação ao mundo real. • Implicação em relação ato de produção

Anexo 3: unidades linguísticas analisadas

As unidades linguísticas analisadas têm como objetivo identificar os tipos discursivos. Neste sentido, a sua análise orienta-se no sentido de identificar os elementos que distinguem os eixo temporal (conjunção e disjunção) do eixo atorial (implicação e autonomia). Utilizámos as etiquetas de anotação [55]do *Sketch Engine*.

1. Densidade Verbal, Nominal, Pronominal, Adverbial: [tag="V.*"]; [tag="N.*"]; [tag="P.*"]; [tag="R.*"]

a. V. - *verbo*

b. N. - *nome*

c. P. - *pronome*

d. R. - *advérbios*

2. Caracterização morfossintática dos verbos (informação de pessoa).

a. Pessoa: [tag="V.*1.*"]; [tag="V.*2.*"]; [tag="V.*3.*"]

1 - *primeira pessoa (singular ou plural)*

2 - *segunda pessoa (singular ou plural)*

3 - *terceira pessoa (singular ou plural)*

b. Tempos verbais do Modo Indicativo

P. - *presente* : [tag="VMIP.*"]

I. - *imperfeito:* [tag="VMII.*"]

F. - *futuro*: [tag="VMIF.*"]

S. - *passado* [tag="VMIS.*"]

[55] Disponível em: https://www.sketchengine.eu/portuguese-freeling-part-of-speech-tagset/

C.- *condicional* [tag="VMIC.*"]

M.- *Pretérito mais-que-perfeito* [lemma="ter" & tag="VMII.*"] [tag="VMP.*"]

PPC. - *Pretérito Perfeito Composto* [lemma="ter"& tag="VMIP.*"]+[tag="V.*"].

c. Modo

I. - indicativo [tag="VMI.*"]

S. - conjuntivo [tag="VMS.*"]

M. - imperativo [tag="VMM.*"]

P. - particípio passado [tag="VM.*] [tag="VMP.*"] (só quando precedido de auxiliar)

G. - gerúndio [tag="VMG.*"]

N. - infinitivo [tag="VMN.*"]

3. Pronomes (Tabela 1, parte **3.3.1** para os pronomes com função deítica)

a. sujeito[word="eu|tu|nós"]

eu

tu

nós

b. objeto direto [word="me|te "]

me

te

nos

c. complemento oblíquo: [tag="S.*"]+[word="mim|comigo|ti|contigo|nós|connosco"]

mim

comigo

ti

contigo

nós

connosco

d. agente da passiva, introduzidos pela preposição "por": [word="por"] [lemma="mim|ti|nós"]

mim

ti

nós

4. Advérbios

a. [word="aqui|aí|ali|cá|lá|acolá|além"]

aqui

aí

ali

cá

lá

acolá

além

b. de tempo (com função deítica): [word="agora|amanhã|hoje|ontem|anteontem"]

agora

amanhã

hoje

ontem

anteontem

b1. Locuções temporais. Entendemos por locuções temporais os grupos preposicionais com referências temporais modificados por adjetivos com valor referencial. (Hagège et al., 2010)

[lemma="dia|hora|semana|ano|década"] [lemma="passar"]

[lemma="próximo"] [lemma="dia|hora|semana|ano|década"]

Anexo 4: tabela de frequência relativa dos *subcorpora.*

Frequências relativas dos *subcorpora* (em percentagem)

$$Frequência\ Relativa = \left(\frac{Valor\ absoluto}{Total\ de\ palavras}\right) X\ 100$$

Tabela com as frequências relativas dos três *subcorpora* (CETEM, COMENTA2, COMJUR):

Variável	CETEM (Freq. Relativa %)	COMENTA2 (Freq. Relativa %)	COMJUR (Freq. Relativa %)
Verbo	14,13	14,24	13,57
Nome	33,24	31,42	31,9
Pronome	3,74	5,3	5,25
Advérbio	5,09	5,52	5,24
1ª pessoa	0,89	1,21	0,83
2ª pessoa	0,12	0,19	0
3ª pessoa	7,62	6,99	6,45
Indicativo Presente	4,35	5,76	4,91
Indicativo Imperfeito	0,59	0,43	0,04
Indicativo Futuro Simples	0,43	0,24	0,49
Indicativo Pret. Perf. Simples	2,19	0,9	0,63
Condicional	0,17	0,22	0,16
Mais-que-Perfeito Indicativo	0,04	0,02	0,02
Pret. Perf. Composto Indicativo	0,07	0,04	0,07
Modo Indicativo	7,81	7,69	6,25
Modo Conjuntivo	0,58	0,68	0,8
Modo Imperativo	0	0,01	0
Particípio Passado	2,4	2,42	2,97

Gerúndio	0,42	0,62	0,76
Infinitivo	2,91	2,63	2,79
Pronomes Sujeito	0,05	0,12	0,04
Pronomes obj. direto	0,03	0,07	0
Pronomes oblíquos	0,02	0,06	0,04
Pronomes agente da passiva	0	0	0,07
Advérbios de Lugar	0,13	0,19	0,25
Advérbios de Tempo	0,29	0,08	0,09

Anexo 5: tabela de frequência dos Tipos Discursivos

ID	DI	DT	N	RI
102	4	7	3	0
104	4	13	1	7
702	5	3	0	1
703	4	4	0	0
704	1	1	1	1
705	4	1	3	1
706	5	2	1	0
707	3	4	0	2
708	1	1	3	2
712	1	1	0	0
717	0	3	0	3
718	0	2	0	2
719	0	2	0	1
720	0	3	3	0
721	1	2	2	0
722	1	2	0	1
723	2	3	0	0
724	2	3	1	0
725	2	4	2	0
726	1	4	2	0
727	1	1	0	0
728	2	3	1	0
729	3	3	1	0

730	6	9	0	5
731	3	10	8	3
732	0	11	0	5
733	4	13	4	3
734	0	3	5	4
735	0	7	4	0
736	3	4	0	2
737	1	10	3	3
738	5	6	0	0
739	11	16	3	1
740	3	5	0	1
742	1	4	1	0
743	5	5	5	1
744	0	4	0	1
745	6	10	1	0
746	0	3	0	0
747	0	9	2	1
748	2	5	2	2
750	0	2	0	0
751	2	5	4	1
752	0	5	0	3
753	0	1	0	0
754	0	2	0	0
755	0	5	3	1
756	0	11	1	0
757	0	2	0	0

758	1	16	2	3
759	0	6	2	2
760	3	12	2	5
761	0	7	0	5
762	2	4	0	1
764	1	7	7	0
765	1	4	4	0
766	0	9	1	0
767	1	6	3	2
768	0	4	3	0
769	1	12	2	0
770	0	6	0	0
771	0	6	3	0
772	4	5	0	0
773	2	13	1	0
774	2	7	0	0
775	0	1	0	0
776	1	3	1	0
777	0	9	7	0
778	1	3	1	0
779	1	4	2	0
781	7	13	1	3
782	2	8	4	0
783	1	5	0	0
784	9	8	2	4
785	2	9	2	1

789	0	2	2	0
790	0	3	2	1
791	0	1	1	0
792	4	5	0	0
793	3	4	1	0
794	1	1	1	0
795	1	5	2	2

Anexo 6: dados estatísticos do modelo GLM

As contagens e dados estatísticos encontram-se descritos nos gráficos e tabelas seguintes. Nas tabelas, a coluna “Value” identifica a variável em análise, a coluna “Count” contém os números absolutos e a coluna “Percentage” a percentagem que o valor representa no total.

1. Tipos Discursivos (TD)

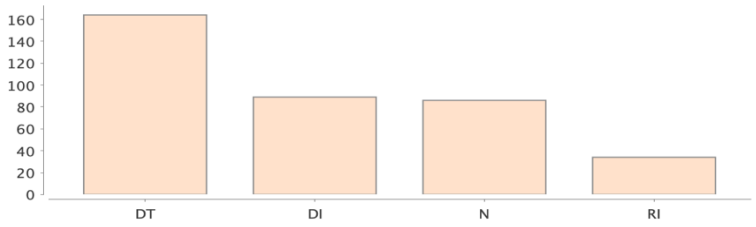
TD

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 1.07%
Stability: 43.97%
Valid: 54.96%

Top Values



4 Distinct Values:

Value	Count	Percentage
DT	164	43.97%
DI	89	23.86%
N	86	23.06%
RI	34	9.12%

2. Atividade

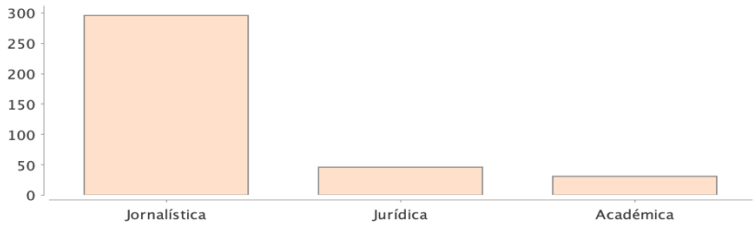
Atividade

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.80%
Stability: 79.36%
Valid: 19.84%

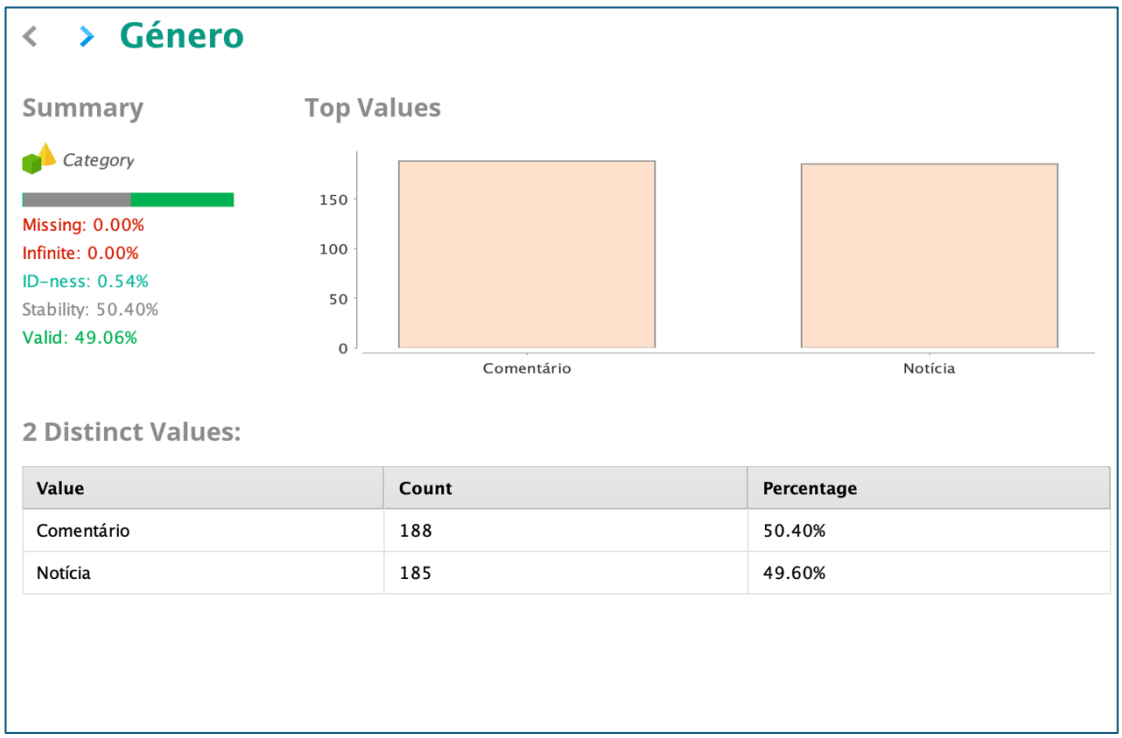
Top Values



3 Distinct Values:

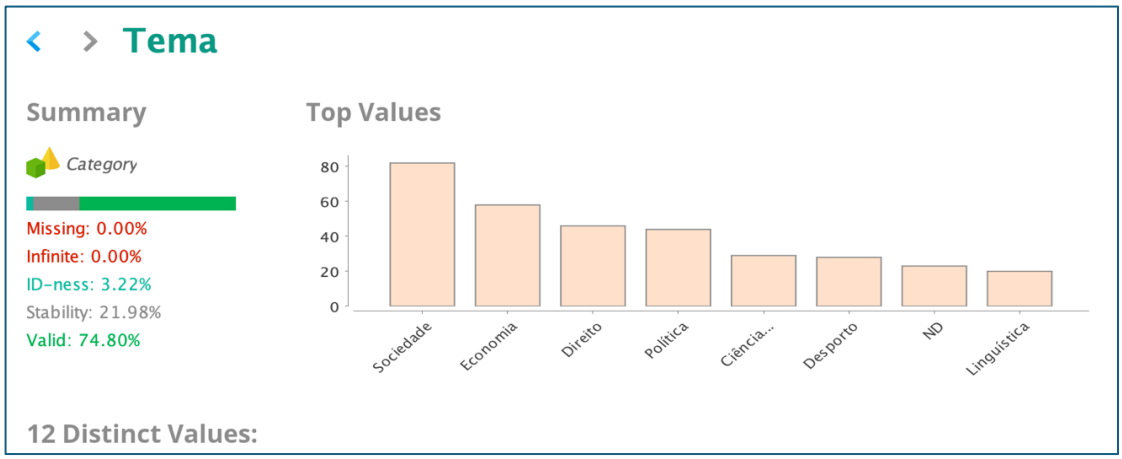
Value	Count	Percentage
Jornalística	296	79.36%
Jurídica	46	12.33%
Académica	31	8.31%

3. Género



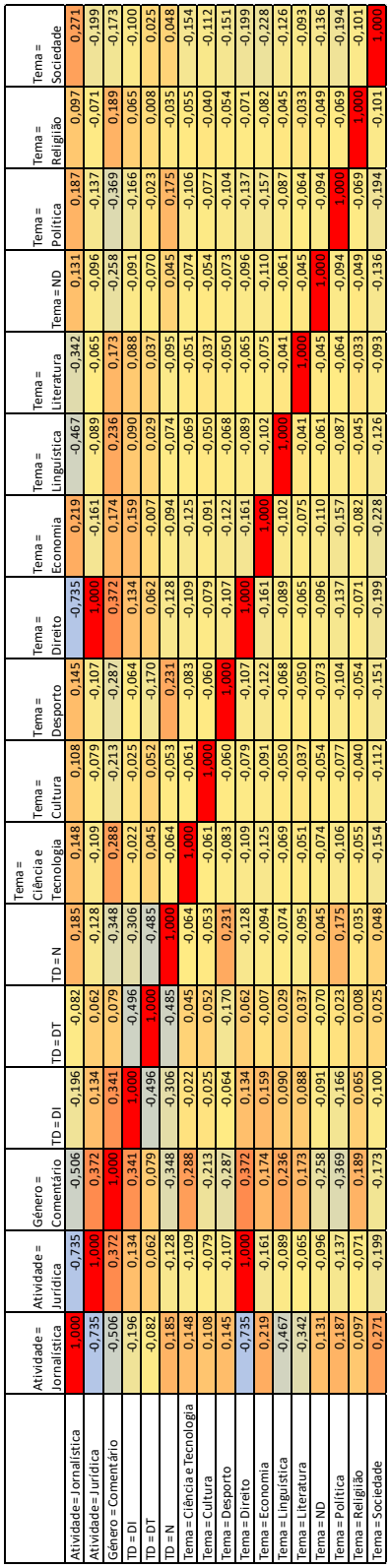
Value	Count	Percentage
Comentário	188	50.40%
Notícia	185	49.60%

4. Tema



Value	Count	Pecentage
Sociedade	82	21,98%
Religião	13	3,49%
Política	44	11,80%
Opinião	3	0,80%
ND	23	6,17%
Literatura	11	2,95%
Linguística	20	5,36%
Economia	58	15,55%
Direito	46	12,33%
Desporto	28	7,51%
Cultura	16	4,29%
Ciência e Tecnologia	29	7,77%

Anexo 7: matriz de correlação das variáveis do modelo



	Atividade = Jornalística	Atividade = Jurídica	Género = Comentário	TD = DI	TD = DT	TD = N	Tema = Ciência e Tecnologia	Tema = Cultura	Tema = Desporto	Tema = Direito	Tema = Economia	Tema = Linguística	Tema = Literatura	Tema = ND	Tema = Política	Tema = Religião	Tema = Sociedade
Atividade = Jornalística	1,000	-0,735	-0,506	-0,196	-0,082	0,185	0,148	0,108	0,145	-0,735	0,219	-0,467	-0,342	0,131	0,187	0,097	0,271
Atividade = Jurídica	-0,735	1,000	0,372	0,134	0,062	-0,128	-0,109	-0,079	-0,107	1,000	-0,161	-0,089	-0,065	-0,096	-0,137	-0,071	-0,199
Género = Comentário	-0,506	0,372	1,000	0,341	0,079	-0,348	0,288	-0,213	-0,287	0,372	0,174	0,236	0,173	-0,258	-0,369	0,189	-0,173
TD = DI	-0,196	0,134	0,341	1,000	-0,496	-0,306	-0,022	-0,025	-0,064	0,134	0,159	0,090	0,088	-0,091	-0,166	0,065	-0,100
TD = DT	-0,082	0,062	0,079	-0,496	1,000	-0,485	0,045	0,052	-0,170	0,062	-0,007	0,029	0,037	-0,070	-0,023	0,008	0,025
TD = N	0,185	-0,128	-0,348	-0,306	-0,485	1,000	-0,064	-0,053	0,231	-0,128	-0,094	-0,074	-0,095	0,045	0,175	-0,035	0,048
Tema = Ciência e Tecnologia	0,148	-0,109	0,288	-0,022	0,045	-0,064	1,000	-0,061	-0,083	-0,109	-0,125	-0,069	-0,051	-0,074	-0,106	-0,055	-0,154
Tema = Cultura	0,108	-0,079	-0,213	-0,025	0,052	-0,053	-0,061	1,000	-0,060	-0,079	-0,091	-0,050	-0,037	-0,054	-0,077	-0,040	-0,112
Tema = Desporto	0,145	-0,107	-0,287	-0,064	-0,170	0,231	-0,083	-0,060	1,000	-0,107	-0,122	-0,068	-0,050	-0,073	-0,104	-0,054	-0,151
Tema = Direito	-0,735	1,000	0,372	0,134	0,062	-0,128	-0,109	-0,079	-0,107	1,000	-0,161	-0,089	-0,065	-0,096	-0,137	-0,071	-0,199
Tema = Economia	0,219	-0,161	0,174	0,159	-0,007	-0,094	-0,125	-0,091	-0,122	-0,161	1,000	-0,102	-0,075	-0,110	-0,157	-0,082	-0,228
Tema = Linguística	-0,467	-0,089	0,236	0,090	0,029	-0,074	-0,069	-0,050	-0,068	-0,089	-0,102	1,000	-0,041	-0,061	-0,087	-0,045	-0,126
Tema = Literatura	-0,342	-0,065	0,173	0,088	0,037	-0,095	-0,051	-0,037	-0,050	-0,065	-0,075	-0,041	1,000	-0,045	-0,064	-0,033	-0,093
Tema = ND	0,131	-0,096	-0,258	-0,091	-0,070	0,045	-0,074	-0,054	-0,073	-0,096	-0,110	-0,061	-0,045	1,000	-0,094	-0,069	-0,136
Tema = Política	0,187	-0,137	-0,369	-0,166	-0,023	0,175	-0,106	-0,077	-0,104	-0,137	-0,157	-0,087	-0,064	-0,094	1,000	-0,069	-0,194
Tema = Religião	0,097	-0,071	0,189	0,065	0,008	-0,035	-0,055	-0,040	-0,054	-0,071	-0,082	-0,045	-0,033	-0,049	-0,069	1,000	-0,101
Tema = Sociedade	0,271	-0,199	-0,173	-0,100	0,025	0,048	-0,154	-0,112	-0,151	-0,199	-0,228	-0,126	-0,093	-0,136	-0,194	-0,101	1,000

Figura 22: Matriz de Correlações dos Atributos.

Anexo 8: tabela de previsões do modelo

Tabela de Previsões

Tema	Atividade	TD	Género	confidence (Notícia)	confidence (Comentário)	prediction (Género)	cost
Sociedade	Jornalística	DT	Comentário	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	DI	Comentário	0,4	0,6	Comentário	0,2
Sociedade	Jornalística	DI	Comentário	0,4	0,6	Comentário	0,2
Sociedade	Jornalística	DI	Comentário	0,4	0,6	Comentário	0,2
Sociedade	Jornalística	DT	Comentário	0,7	0,3	Notícia	0,4
Economia	Jornalística	DT	Comentário	0,3	0,7	Comentário	0,3
Economia	Jornalística	RI	Comentário	0,4	0,6	Comentário	0,3
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Sociedade	Jornalística	RI	Comentário	0,7	0,3	Notícia	0,4

Sociedade	Jornalística	N	Comentário	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	DI	Comentário	0,4	0,6	Comentário	0,2
Sociedade	Jornalística	N	Comentário	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	DI	Comentário	0,4	0,6	Comentário	0,2
Sociedade	Jornalística	RI	Comentário	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	DT	Comentário	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	DT	Comentário	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	RI	Comentário	0,7	0,3	Notícia	0,4
Ciência e Tecnologia	Jornalística	DT	Comentário	0,2	0,8	Comentário	0,6
Ciência e Tecnologia	Jornalística	RI	Comentário	0,2	0,8	Comentário	0,6
Ciência e Tecnologia	Jornalística	N	Comentário	0,5	0,5	Comentário	0,0
Ciência e Tecnologia	Jornalística	DT	Comentário	0,2	0,8	Comentário	0,6
Ciência e Tecnologia	Jornalística	DT	Comentário	0,2	0,8	Comentário	0,6
Ciência e Tecnologia	Jornalística	DT	Comentário	0,2	0,8	Comentário	0,6
Ciência e Tecnologia	Jornalística	DT	Comentário	0,2	0,8	Comentário	0,6
Ciência e Tecnologia	Jornalística	DT	Comentário	0,2	0,8	Comentário	0,6

Ciência e Tecnologia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,7
Ciência e Tecnologia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,7
Ciência e Tecnologia	Jornalística	N	Comentário	0,5	0,5	Comentário	0,0
Ciência e Tecnologia	Jornalística	DT	Comentário	0,2	0,8	Comentário	0,6
Ciência e Tecnologia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,7
Religiião	Jornalística	DT	Comentário	0,3	0,7	Comentário	0,3
Religiião	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Religiião	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Religiião	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Literatura	Académica	DI	Comentário	0,2	0,8	Comentário	0,6
Literatura	Académica	DT	Comentário	0,3	0,7	Comentário	0,4
Literatura	Académica	DT	Comentário	0,3	0,7	Comentário	0,4
Direito	Jurídica	DT	Comentário	0,2	0,8	Comentário	0,5
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	N	Comentário	0,3	0,7	Comentário	0,3
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7

Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	DT	Comentário	0,2	0,8	Comentário	0,5
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	DT	Comentário	0,2	0,8	Comentário	0,5
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	DT	Comentário	0,2	0,8	Comentário	0,5
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	DI	Comentário	0,2	0,8	Comentário	0,7
Direito	Jurídica	DT	Comentário	0,2	0,8	Comentário	0,5
Linguística	Académica	N	Comentário	0,4	0,6	Comentário	0,3
Linguística	Académica	DI	Comentário	0,2	0,8	Comentário	0,6
Linguística	Académica	DT	Comentário	0,3	0,7	Comentário	0,4
Linguística	Académica	DI	Comentário	0,2	0,8	Comentário	0,6

Linguística	Académica	DT	Comentário	0,3	0,7	Comentário	0,4
Linguística	Académica	DI	Comentário	0,2	0,8	Comentário	0,6
Linguística	Académica	DT	Comentário	0,3	0,7	Comentário	0,4
Economia	Jornalística	DT	Comentário	0,3	0,7	Comentário	0,3
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Economia	Jornalística	DT	Comentário	0,3	0,7	Comentário	0,3
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Economia	Jornalística	N	Comentário	0,7	0,3	Notícia	0,4
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Economia	Jornalística	DT	Comentário	0,3	0,7	Comentário	0,3
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Economia	Jornalística	DT	Comentário	0,3	0,7	Comentário	0,3
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6

Economia	Jornalística	DI	Comentário	0,2	0,8	Comentário	0,6
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Política	Jornalística	DI	Notícia	0,7	0,3	Notícia	0,4
Política	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Cultura	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Cultura	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Economia	Jornalística	N	Notícia	0,7	0,3	Notícia	0,4
Economia	Jornalística	DT	Notícia	0,3	0,7	Comentário	0,3
Economia	Jornalística	DI	Notícia	0,2	0,8	Comentário	0,6
Desporto	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Política	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Política	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6

Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Política	Jornalística	DI	Notícia	0,7	0,3	Notícia	0,4
ND	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Desporto	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
ND	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
ND	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Cultura	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Economia	Jornalística	N	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6
Economia	Jornalística	DI	Notícia	0,2	0,8	Comentário	0,6
Política	Jornalística	RI	Notícia	0,8	0,2	Notícia	0,6
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Política	Jornalística	RI	Notícia	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6

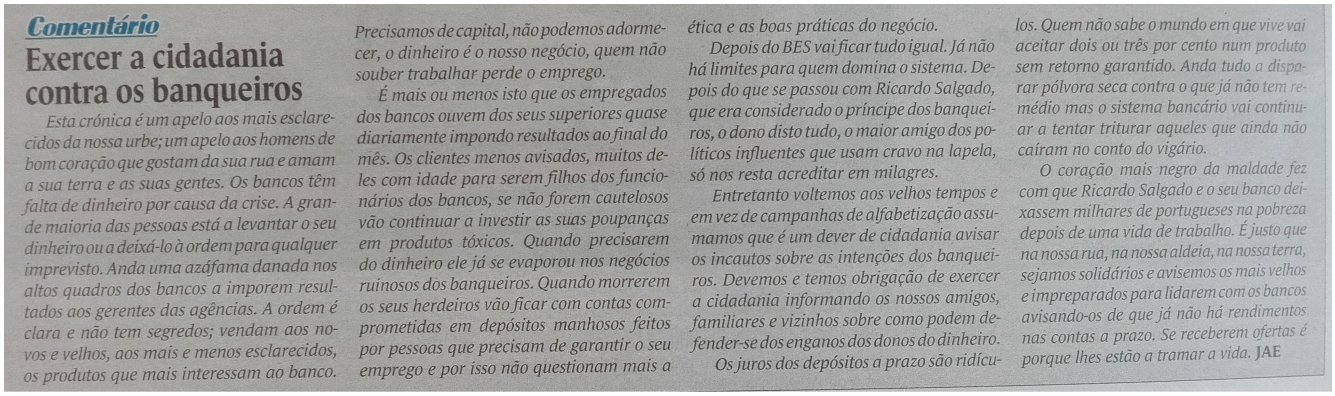
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	RI	Notícia	0,7	0,3	Notícia	0,4
Desporto	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
ND	Jornalística	RI	Notícia	0,8	0,2	Notícia	0,6
Desporto	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Economia	Jornalística	N	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Cultura	Jornalística	RI	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	RI	Notícia	0,7	0,3	Notícia	0,4
Política	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
ND	Jornalística	RI	Notícia	0,8	0,2	Notícia	0,6
ND	Jornalística	DI	Notícia	0,6	0,4	Notícia	0,3

Desporto	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Política	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Cultura	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
ND	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
ND	Jornalística	DI	Notícia	0,6	0,4	Notícia	0,3
Cultura	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Desporto	Jornalística	RI	Notícia	0,8	0,2	Notícia	0,6
Desporto	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Política	Jornalística	RI	Notícia	0,8	0,2	Notícia	0,6
Cultura	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6

Desporto	Jornalística	DI	Notícia	0,7	0,3	Notícia	0,4
Economia	Jornalística	DT	Notícia	0,3	0,7	Comentário	0,3
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
ND	Jornalística	N	Notícia	0,8	0,2	Notícia	0,7
Sociedade	Jornalística	N	Notícia	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Cultura	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4
Política	Jornalística	DT	Notícia	0,8	0,2	Notícia	0,6
Sociedade	Jornalística	DT	Notícia	0,7	0,3	Notícia	0,4

Anexo 9: anotação do texto 702 e 703 em XML

Texto 702



Comentário

Exercer a cidadania contra os banqueiros

Esta crónica é um apelo aos mais esclarecidos da nossa urbe; um apelo aos homens de bom coração que gostam da sua rua e amam a sua terra e as suas gentes. Os bancos têm falta de dinheiro por causa da crise. A grande maioria das pessoas está a levantar o seu dinheiro ou a deixá-lo à ordem para qualquer imprevisto. Anda uma azáfama danada nos altos quadros dos bancos a imporem resultados aos gerentes das agências. A ordem é clara e não tem segredos; vendam aos novos e velhos, aos mais e menos esclarecidos, os produtos que mais interessam ao banco. Precisamos de capital, não podemos adormecer, o dinheiro é o nosso negócio, quem não souber trabalhar perde o emprego.

É mais ou menos isto que os empregados dos bancos ouvem dos seus superiores quase diariamente impondo resultados ao final do mês. Os clientes menos avisados, muitos deles com idade para serem filhos dos funcionários dos bancos, se não forem cautelosos vão continuar a investir as suas poupanças em produtos tóxicos. Quando precisarem do dinheiro ele já se evaporou nos negócios ruinosos dos banqueiros. Quando morrerem os seus herdeiros vão ficar com contas comprometidas em depósitos manhosos feitos por pessoas que precisam de garantir o seu emprego e por isso não questionam mais a ética e as boas práticas do negócio.

Depois do BES vai ficar tudo igual. Já não há limites para quem domina o sistema. Depois do que se passou com Ricardo Salgado, que era considerado o príncipe dos banqueiros, o dono disto tudo, o maior amigo dos políticos influentes que usam cravo na lapela, só nos resta acreditar em milagres.

Entretanto voltemos aos velhos tempos e em vez de campanhas de alfabetização assumamos que é um dever de cidadania avisar os incautos sobre as intenções dos banqueiros. Devemos e temos obrigação de exercer a cidadania informando os nossos amigos, familiares e vizinhos sobre como podem defender-se dos enganos dos donos do dinheiro.

Os juros dos depósitos a prazo são ridículos. Quem não sabe o mundo em que vive vai aceitar dois ou três por cento num produto sem retorno garantido. Anda tudo a disparar pólvora seca contra o que já não tem remédio mas o sistema bancário vai continuar a tentar triturar aqueles que ainda não caíram no conto do vigário.

O coração mais negro da maldade fez com que Ricardo Salgado e o seu banco deixassem milhares de portugueses na pobreza depois de uma vida de trabalho. É justo que na nossa rua, na nossa aldeia, na nossa terra, sejamos solidários e avisemos os mais velhos e impreparados para lidarem com os bancos avisando-os de que já não há rendimentos nas contas a prazo. Se receberem ofertas é porque lhes estão a tramar a vida. **JAE**

```
<Comentário>
      <T ID="702">
            <DI>
                  "Comentário Exercer a cidadania contra os banqueirosEsta
crónica é um apelo aos mais esclarecidos da nossa urbe; um apelo aos homens de
bom coração que gostam da sua rua e amam a sua terra e as suas gentes. </DI> "
            <DT>
                  Os bancos têm falta de dinheiro por causa da crise. </DT>
            <DI>
                  A grande maioria das pessoas está a levantar o seu
dinheiro ou a deixá-lo à ordem para qualquer imprevisto. Anda uma azáfama
danada nos altos quadros dos bancos a imporem resultados aos gerentes das
agências. A ordem é clara e não tem segredos, vendam aos novos e velhos, aos
mais e menos esclarecidos, os produtos que mais interessam ao banco.</DI>
            <RI>
                  Precisamos de capital, não podemos adormecer, o dinheiro é
o nosso negócio, quem não souber trabalhar perde o emprego. é mais ou menos
isto que os empregados dos bancos ouvem dos seus superiores quase diariamente
impondo resultados ao final do mês. Os clientes menos avisados, muitos deles
com idade para serem filhos dos funcionários dos bancos, se não forem
cautelosos vão continuar a investir as suas poupanças em produtos
tóxicos.</RI>
            <DI>
                  Quando precisarem do dinheiro ele já se evaporou nos
negócios ruinosos dos banqueiros. Quando morrerem os seus herdeiros vão ficar
com contas comprometidas em depósitos manhosos feitos por pessoas que precisam
de garantir o seu emprego e por isso não questionam mais a Ética e as boas
práticas do negócio.</DI>
```

<DT>

Depois do BES vai ficar tudo igual. Já não há limites para quem domina o sistema. </DT>

<DI>

Depois do que se passou com Ricardo Salgado, que era considerado o príncipe dos banqueiros, o dono disto tudo, o maior amigo dos políticos influentes que usam cravo na lapela, só nos resta acreditar em milagres. Entretanto voltemos aos velhos tempos e em vez de campanhas de alfabetização assumamos que é um dever de cidadania avisar os incautos sobre as intenções dos banqueiros. Devemos e temos obrigação de exercer a cidadania informando os nossos amigos, familiares e vizinhos sobre como podem defender-se dos enganos dos donos do dinheiro.</DI>

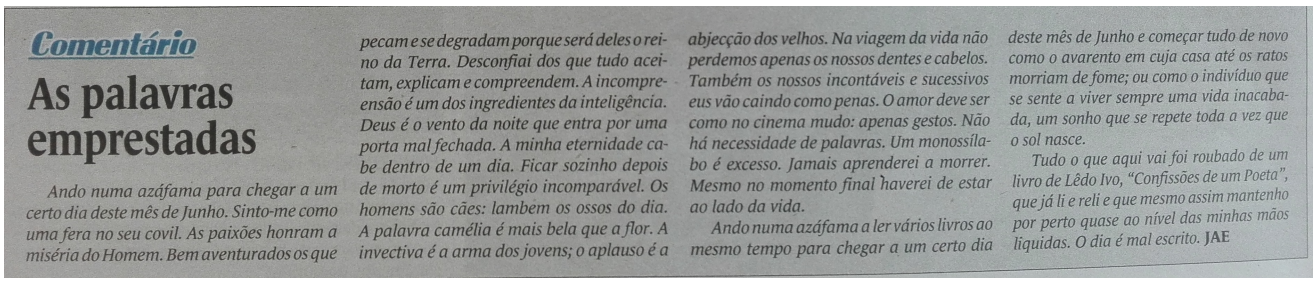
<DT>

Os juros dos depósitos a prazo são ridículos.</DT>

<DI>

Quem não sabe o mundo em que vive vai aceitar dois ou três por cento num produto sem retorno garantido. Anda tudo a disparar pólvora seca contra o que já não tem remédio mas o sistema bancário vai continuar a tentar triturar aqueles que ainda não caíram no conto do vigário. O coração mais negro da maldade fez com que Ricardo Salgado e o seu banco deixassem milhares de portugueses na pobreza depois de uma vida de trabalho. é justo que na nossa rua, na nossa aldeia, na nossa terra, sejamos solidários e avisemos os mais velhos e impreparados para lidarem com os bancos avisando-os de que já não há rendimentos nas contas a prazo. Se receberem ofertas é porque lhes estão a tramar a vida.JAE</DI>

Texto 703



Comentário

As palavras emprestadas

Ando numa azáfama para chegar a um certo dia deste mês de Junho. Sinto-me como uma fera no seu covil. As paixões honram a miséria do Homem. Bem aventurados os que pecam e se degradam porque será deles o reino da Terra. Desconfiai dos que tudo aceitam, explicam e compreendem. A incompreensão é um dos ingredientes da inteligência. Deus é o vento da noite que entra por uma porta mal fechada. A minha eternidade cabe dentro de um dia. Ficar sozinho depois de morto é um privilégio incomparável. Os homens são cães: lambem os ossos do dia. A palavra camélia é mais bela que a flor. A invectiva é a arma dos jovens; o aplauso é a abjecção dos velhos. Na viagem da vida não perdemos apenas os nossos dentes e cabelos. Também os nossos incontáveis e sucessivos eus vão caindo como penas. O amor deve ser como no cinema mudo: apenas gestos. Não há necessidade de palavras. Um monossílabo é excesso. Jamais aprenderei a morrer. Mesmo no momento final haverei de estar ao lado da vida.

Ando numa azáfama a ler vários livros ao mesmo tempo para chegar a um certo dia deste mês de Junho e começar tudo de novo como o avarento em cuja casa até os ratos morriam de fome; ou como o indivíduo que se sente a viver sempre uma vida inacabada, um sonho que se repete toda a vez que o sol nasce.

Tudo o que aqui vai foi roubado de um livro de Lêdo Ivo, "Confissões de um Poeta", que já li e reli e que mesmo assim mantenho por perto quase ao nível das minhas mãos liquidas. O dia é mal escrito. JAE

```
<T ID="703">

            <DI>

                  Comentário As palavras emprestadas Ando numa azáfama para
chegar a um certo dia deste mês de Junho. Sinto-me como uma fera no seu covil.
</DI>

            <DT>

                  As paixões honram a miséria do Homem. Bem aventurados os
que pecam e se degradam porque será deles o reino da Terra.</DT>

            <DI>

                  Desconfia dos que tudo aceitam, explicam e compreendem.
</DI>

            <DT>

                  A incompreensão é um dos ingredientes da inteligência.
Deus é o vento da noite que entra por uma porta mal fechada.</DT>

            <DI>

                  A minha eternidade cabe dentro de um dia.</DI>

            <DT>

                  Ficar sozinho depois de morto é um privilégio
incomparável. Os homens são cães: lambem os ossos do dia. A palavra camélia é
mais bela que a flor. A invectiva é a arma dos jovens, o aplauso é a abjecção
dos velhos. Na viagem da vida não perdemos apenas os nossos dentes e cabelos.
Também os nossos incontáveis e sucessivos eus vão caindo como penas. O amor
deve ser como no cinema mudo: apenas gestos. Não há necessidade de palavras.
Um monossílabo é excesso.</DT>

            <DI>

                  "Jamais aprenderei a morrer. Mesmo no momento final
haverei de estar ao lado da vida. Ando numa azáfama a ler vários livros ao
mesmo tempo para chegar a um certo dia deste mês de Junho e começar tudo de
novo como o avarento em cuja casa até os ratos morriam de fome; ou como o
indivíduo que se sente a viver sempre uma vida inacabada, um sonho que se
repete toda a vez que o sol nasce. Tudo o que aqui vai foi roubado de um livro
```

de Lêdo Ivo, "Confissões de um Poeta", que já li e reli e que mesmo assim mantenho por perto quase ao nível das minhas mãos líquidas.</DI>

<DT>

Anexo 10: *Corpora*

Devido ao tamanho dos *corpora* usados neste trabalho, estes encontram-se disponíveis através da ligação: Corpora.

Anexo 11: *Tagset* para o Português

Part of Speech: adjective

POSITION	ATRIBUTE	VALUES
0	category	**A**:*adjective*
1	type	**O**:*ordinal*; **Q**:*qualificative*; **P**:*possessive*
2	degree	**S**:*superlative*; **V**:*evaluative*
3	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*
4	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
5	possessorpers	**1**:*1*; **2**:*2*; **3**:*3*
6	possessornum	**S**:*singular*; **P**:*plural*; **N**:*invariable*

Part of Speech: conjunction

POSITION	ATRIBUTE	VALUES
0	category	**C**:*conjunction*
1	type	**C**:*coordinating*; **S**:*subordinating*

Part of Speech: determiner

POSITION	ATRIBUTE	VALUES
0	category	**D**:*determiner*

POSITION	ATRIBUTE	VALUES
1	type	**A**:*article*; **D**:*demonstrative*; **E**:*exclamative*; **I**:*indefinite*; **T**:*interrogative*;**N**:*numeral*; **P**:*possessive*
2	person	**1**:*1*; **2**:*2*; **3**:*3*
3	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*
4	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
5	possessor num	**S**:*singular*; **P**:*plural*

Part of Speech: noun

POSITION	ATRIBUTE	VALUES
0	category	**N**:*noun*
1	type	**C**:*common*; **P**:*proper*
2	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*
3	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
4	neclass	**S**:*person*; **G**:*location*; **O**:*organization*; **V**:*other*
5	nesubclass	*Not used*
6	degree	**A**:*augmentative*; **D**:*diminutive*

Part of Speech: pronoun

POSITION	ATRIBUTE	VALUES
0	category	**P**:*pronoun*
1	type	**D**:*demonstrative*; **E**:*exclamative*; **I**:*indefinite*; **T**:*interrogative*; **N**:*numeral*;**P**:*personal*; **R**:*relative*
2	person	**1**:*1*; **2**:*2*; **3**:*3*
3	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*
4	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
5	case	**N**:*nominative*; **A**:*accusative*; **D**:*dative*; **O**:*oblique*
6	polite	**P**:*yes*

Part of Speech: adverb

POSITION	ATRIBUTE	VALUES
0	category	**R**:*adverb*
1	type	**N**:*negative*; **G**:*general*

Part of Speech: adposition

POSITION	ATRIBUTE	VALUES
0	category	**S**:*adposition*
1	type	**P**:*preposition*

Part of Speech: verb

POSITION	ATRIBUTE	VALUES
0	category	**V**:*verb*
1	type	**M**:*main*; **A**:*auxiliary*; **S**:*semiauxiliary*
2	mood	**I**:*indicative*; **S**:*subjunctive*; **M**:*imperative*; **P**:*pastparticiple*; **G**:*gerund*;**N**:*infinitive*
3	tense	**P**:*present*; **I**:*imperfect*; **F**:*future*; **S**:*past*; **C**:*conditional*; **M**:*plusquamperfect*
4	person	**1**:*1*; **2**:*2*; **3**:*3*
5	num	**S**:*singular*; **P**:*plural*
6	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*

Part of Speech: number

POSITION	ATRIBUTE	VALUES
0	category	**Z**:*number*
1	type	**d**:*partitive*; **m**:*currency*; **p**:*ratio*; **u**:*unit*

Part of Speech: date

POSITION	ATRIBUTE	VALUES
0	category	**W**:*date*

Part of Speech: interjection

POSITION	ATRIBUTE	VALUES
0	category	**I**:*interjection*

Non-positional tags

Part of Speech: punctuation

TAG	ATTRIBUTES
Fd	**pos**:*punctuation*; **type**:*colon*
Fc	**pos**:*punctuation*; **type**:*comma*
Flt	**pos**:*punctuation*; **type**:*curlybracket*; **punctenclose**:*close*
Fla	**pos**:*punctuation*; **type**:*curlybracket*; **punctenclose**:*open*
Fs	**pos**:*punctuation*; **type**:*etc*
Fat	**pos**:*punctuation*; **type**:*exclamationmark*; **punctenclose**:*close*
Faa	**pos**:*punctuation*; **type**:*exclamationmark*; **punctenclose**:*open*
Fg	**pos**:*punctuation*; **type**:*hyphen*
Fz	**pos**:*punctuation*; **type**:*other*
Fpt	**pos**:*punctuation*; **type**:*parenthesis*; **punctenclose**:*close*
Fpa	**pos**:*punctuation*; **type**:*parenthesis*; **punctenclose**:*open*
Ft	**pos**:*punctuation*; **type**:*percentage*
Fp	**pos**:*punctuation*; **type**:*period*

TAG	ATTRIBUTES
Fit	**pos**:*punctuation*; **type**:*questionmark*; **punctenclose**:*close*
Fia	**pos**:*punctuation*; **type**:*questionmark*; **punctenclose**:*open*
Fe	**pos**:*punctuation*; **type**:*quotation*
Frc	**pos**:*punctuation*; **type**:*quotation*; **punctenclose**:*close*
Fra	**pos**:*punctuation*; **type**:*quotation*; **punctenclose**:*open*
Fx	**pos**:*punctuation*; **type**:*semicolon*
Fh	**pos**:*punctuation*; **type**:*slash*
Fct	**pos**:*punctuation*; **type**:*squarebracket*; **punctenclose**:*close*
Fca	**pos**:*punctuation*; **type**:*squarebracket*; **punctenclose**:*open*

Fonte: https://freeling-user-manual.readthedocs.io/en/latest/tagsets/tagset-pt/