

**Vocabulário Fundamental da Academia das
Ciências de Lisboa: seleção lexical, alinhamento de
sentidos e codificação**

Maria Leonor Velez Máximo dos Reis

**Dissertação de Mestrado em Ciências da Linguagem com
especialização em Terminologia e Gestão de Informação
de Especialidade**

Versão corrigida e melhorada após a sua defesa pública

abril de 2025

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Ciências da Linguagem, área de especialização em Terminologia e Gestão da Informação de Especialidade, realizada sob a orientação científica da Professora Doutora Raquel Amaro e da Professora Doutora Ana Salgado.

Apoio financeiro através do projeto *iRead4Skills – Intelligent Reading Improvement System for Fundamental and Transversal Skills Development* [101094837] , financiado pela Comissão Europeia através do programa HORIZON, tópico HORIZON-CL2-2022-TRANSFORMATIONS-01-07 (*Conditions for the successful development of skills matched to needs*), no âmbito de uma bolsa de investigação com a duração de um ano.

Apoio financeiro através do projeto *Dicionário da Língua Portuguesa* (DLP), da Academia de Ciências de Lisboa (ACL), em versão digital, no âmbito de uma bolsa de investigação atribuída nos termos do Regulamento n.º 582/2022, publicado no *Diário da República*, 2.ª série, n.º 123, de 28 de junho de 2022, disponível no site da ACL e do Estatuto do Bolseiro de Investigação Científica, aprovado pela Lei n.º 40/2004, para o desempenho do trabalho lexicográfico desenvolvido sob a tutela do Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLLP).

AGRADECIMENTOS

Este trabalho não teria sido possível sem o apoio e a orientação das minhas orientadoras, a quem expresso o meu mais sincero agradecimento.

À professora Raquel Amaro, por ter sido o meu primeiro contacto real com as Ciências da Linguagem. Agradeço a confiança que depositou em mim, as oportunidades dadas e o encorajamento constante.

À professora Ana Salgado, pelo acompanhamento atento e constante, por todo o conhecimento que tão generosa e tranquilamente partilha e pelo esforço que faz em me ouvir, mesmo diante de toda a falta de eloquência que me é característica.

Aos colegas do *iRead4Skills*, especialmente à Sílvia, pelo companheirismo, pelos momentos de aprendizagem e pela paciência. Esta dissertação não seria possível sem o trabalho de todos.

À Leonor, colega do ILLLP, pelas dúvidas partilhadas, pelas conversas, pelo apoio e pelos risos ao longo deste percurso.

Agradeço aos meus pais e ao meu irmão pelo apoio constante.

Por fim, agradeço a toda minha família pelo apoio geral ao longo desta fase.

VOCABULÁRIO FUNDAMENTAL DA ACADEMIA DAS CIÊNCIAS DE LISBOA: SELEÇÃO LEXICAL, ALINHAMENTO DE SENTIDOS E CODIFICAÇÃO

MARIA LEONOR REIS

RESUMO

PALAVRAS-CHAVE: lexicologia, vocabulário fundamental, *corpus*, alinhamento de sentidos

A presente dissertação tem como objetivo principal a elaboração de um vocabulário fundamental do português atualizado, digital, anotado e codificado segundo a especificação TEI Lex-0 (Tasovac *et al.*, 2018). Diferenciando-se das abordagens tradicionais, geralmente orientadas para o ensino de português como língua não materna, propõe-se aqui um vocabulário destinado a falantes nativos, com organização por níveis de complexidade e com tratamento semântico explícito das unidades lexicais, em vez de uma lista de lemas não desambiguados ou listagens de vocabulário/unidades lexicais orientadas para o ensino de língua não materna.

Este trabalho faz uso dos léxicos organizados por nível de complexidade, desenvolvidos no âmbito do projeto *iRead4Skills – Intelligent Reading Improvement System for Fundamental and Transversal Skills Development*. Esses léxicos foram extraídos de um *corpus* diversificado e atual, composto por textos provenientes de diferentes domínios e contextos comunicativos. A integração num recurso lexicográfico de referência, como o *Dicionário da Língua Portuguesa* (DLP) da Academia das Ciências de Lisboa, assegura uma maior acessibilidade e usabilidade da ferramenta proposta.

Nesse sentido, foi desenvolvida uma metodologia específica que permite identificar as unidades lexicais relevantes e definir unidades lexicais representativas do vocabulário, procedendo ao seu alinhamento com os sentidos correspondentes no DLP e à criação de metadados estruturados visando à codificação e interoperabilidade.

O fluxo de trabalho seguiu quatro etapas principais: (i) a criação e organização dos *subcorpora* por nível de complexidade; (ii) identificação das unidades lexicais relevantes e sua definição; (iii) alinhamento de unidades lexicais definidas e selecionadas com os sentidos correspondentes no DLP; e (iv) criação dos metadados necessários para codificação das unidades lexicais.

Este Vocabulário Fundamental em formato digital promove uma descrição mais precisa e atualizada da língua considerando diferentes níveis de complexidade, ajustada às necessidades de falantes nativos e tendo aplicações diretas em contextos educativos, tecnológicos e sociopolíticos, o que abre novas possibilidades de investigação em linguística e ensino de línguas.

FUNDAMENTAL VOCABULARY OF THE ACADEMIA DAS CIÊNCIAS DE LISBOA: LEXICAL SELECTION, SENSE ALIGNMENT AND ENCODING

MARIA LEONOR REIS

ABSTRACT

KEYWORDS: lexicology, fundamental vocabulary, corpus, word sense alignment

This dissertation's main objective is to develop an updated fundamental vocabulary of Portuguese, annotated and encoded according to the TEI Lex-0 specification (Tasovac *et al.*, 2018). Unlike traditional approaches, which are generally oriented toward teaching Portuguese as a non-native language, this study proposes a vocabulary intended for native speakers, organized by levels of complexity and with explicit semantic treatment of lexical units, rather than lists of non-disambiguated lemmas or vocabulary/lexical unit listings designed for foreign language teaching.

This study leverages on the lexicons organized by complexity level, developed within the framework of the iRead4Skills project – Intelligent Reading Improvement System for Fundamental and Transversal Skills Development. These lexicons were extracted from a diverse and up-to-date corpus composed of texts from various domains and communicative contexts. Their integration into a reference lexicographic resource, such as the *Dicionário da Língua Portuguesa* (DLP) of the Academia das Ciências de Lisboa, ensures greater accessibility and usability of the proposed tool.

To this end, a specific methodology was developed to identify the relevant units and define representative lexical units of the vocabulary, aligning them with their corresponding senses in the DLP and creating structured metadata for encoding and interoperability purposes.

The workflow followed four main steps: (i) creation and organization of subcorpora by complexity level; (ii) identification of the relevant lexical units and their definition; (iii) alignment of selected and defined lexical units with their corresponding senses in the DLP; and (iv) creation of the necessary metadata for encoding the lexical units.

This digital Fundamental Vocabulary provides a more precise and updated description of the Portuguese language considering different complexity levels, and tailored to the needs of native speakers. It has direct applications in educational, technological, and sociopolitical contexts, opening new possibilities for research in linguistics and language teaching.

ÍNDICE

LISTA DE ABREVIATURAS	7
1. INTRODUÇÃO	1
1.1. Integração do trabalho em projetos de investigação	1
1.2. Contextualização da proposta	1
1.3. Objetivos	3
1.4. Estrutura da dissertação	5
2. ENQUADRAMENTO TEÓRICO	6
2.1. Complexidade textual	6
2.1.1. Níveis de complexidade do <i>iRead4Skills</i>	7
2.2. Léxico e complexidade	9
2.2.1. Métodos para a definição de léxicos de complexidade	10
2.2.2. Léxico vs. vocabulário	13
2.2.3. Vocabulários fundamentais: métodos, objetivos e limitações	13
2.3. Alinhamento e integração de conteúdo lexical	15
2.3.1. Alinhamento de sentidos	15
2.3.2. Desafios e metodologias no alinhamento de sentidos	16
2.3.3. Integração e disponibilização de informação lexical	18
2.4. Codificação	19
2.4.1. Normas para o trabalho lexicográfico	21
2.4.2. TEI – <i>Text Encoding Initiative</i>	22
2.4.3. Estratégias de codificação do DLP	24
3. RECURSOS PARA O DESENVOLVIMENTO DO VOCABULÁRIO FUNDAMENTAL	26
3.1. <i>iRead4Skills Dataset 1</i> (v. 2.0): <i>corpora</i> por nível de complexidade	26
3.1.1. <i>Design</i> do <i>corpus</i> : tipologia textual	27
3.1.2. Critérios de seleção e amostragem	28
3.1.3. Validação dos dados	29

3.1.4. Codificação dos metadados	30
3.1.5. Caracterização final do <i>corpus</i>	32
3.2. Léxicos por níveis de complexidade	34
4. DESENVOLVIMENTO DO VOCABULÁRIO FUNDAMENTAL	36
4.1. Criação e organização dos <i>subcorpora</i> por nível de complexidade	37
4.2. Seleção dos lemas	38
4.2.1. Identificação de ULs em <i>corpora</i>	39
4.3. Tratamento das ULs: análise e organização da informação	41
4.3.1. Análise do tratamento lexicográfico no DLP	41
4.3.2. Análise das ULs e elaboração de definições	43
4.4. Alinhamento de sentidos: processo e dificuldades	45
4.5. Metadados para codificação das ULs	52
5. INTEGRAÇÃO DO VOCABULÁRIO FUNDAMENTAL NO DLP	55
5.1. Proposta de <i>design</i> gráfico	55
5.2. Implementação do esquema de codificação	56
6. CONSIDERAÇÕES FINAIS E TRABALHO FUTURO	60
REFERÊNCIAS BIBLIOGRÁFICAS	62
LISTA DE FIGURAS E TABELAS	72
ANEXOS	74
Anexo I – Descritores linguísticos e extralinguísticos por níveis de complexidade	74
Anexo II – Tipologia textual completa	77
Anexo III – Distribuição total do número de textos e <i>tokens</i> por propósito, meio, tipo de documento e nível de complexidade	80
Anexo IV – Identificação de ULs e alinhamento de sentidos	81
Anexo V – Entrada “bater” antes da revisão	98
Anexo VI – Entrada “bater” após revisão e edição	104

LISTA DE ABREVIATURAS

ACL – Academia das Ciências de Lisboa

DLP – *Dicionário da Língua Portuguesa*

DLPC – *Dicionário da Língua Portuguesa Contemporânea*

EFA – Educação e Formação de Adultos

ILLP – Instituto de Lexicologia e Lexicografia da Língua Portuguesa

MWE – *Multiword expression*

PLE – Português língua estrangeira

PLN – Processamento da língua natural

POS – *Part-of-speech*

TEI – *Text Encoding Initiative*

UL – Unidade lexical

WSA – *Word sense alignment*

1. INTRODUÇÃO

1.1. Integração do trabalho em projetos de investigação

A presente dissertação surge no seguimento do trabalho desenvolvido em torno de dois projetos distintos: 1) o projeto *iRead4Skills – Intelligent Reading Improvement System for Fundamental and Transversal Skills Development*¹, financiado pela Comissão Europeia através do programa HORIZON, no âmbito de uma bolsa de investigação com a duração de um ano; e 2) o projeto *Dicionário da Língua Portuguesa*² (DLP), da Academia de Ciências de Lisboa (ACL), em versão digital, no âmbito de uma bolsa de investigação atribuída nos termos do Regulamento de Bolsas de Investigação da ACL, para o desempenho do trabalho lexicográfico desenvolvido sob a tutela do Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLP).

1.2. Contextualização da proposta

A capacidade de leitura desempenha um papel fundamental na aquisição de novos conhecimentos, no aprofundamento dos já existentes e na compreensão do mundo ao nosso redor. No entanto, nem todos os indivíduos têm as mesmas oportunidades de desenvolver essa competência de forma plena.

Pessoas com baixa literacia³ enfrentam diversas dificuldades no que concerne à aquisição e à manutenção de capacidades necessárias para se enquadrarem e permanecerem num mercado de trabalho em constante evolução (OECD, 2019). Em contrapartida, aquelas com níveis de literacia mais elevados apresentam diversos benefícios nos seus postos de trabalho, com possibilidades de progressão de carreira a uma remuneração salarial mais elevada (*ibidem*).

¹ *iRead4Skills – Intelligent Reading Improvement System for Fundamental and Transversal Skills Development*. Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07 – Conditions for the successful development of skills matched to needs. <https://iread4skills.com/>

² <https://dicionario.acad-ciencias.pt/> (consultado a 02/06/2025)

³ A noção de literacia é perspectivada como a capacidade que cada indivíduo apresenta em usar as capacidades de leitura (e de escrita) de forma a participar na sociedade envolvente, impactando as suas vidas social e profissional (OECD, 2024; UNESCO, 2020). É, assim, uma prática social que opera num *continuum*, variando entre indivíduos (Monteiro *et al.*, 2024).

Neste sentido, são cada vez mais relevantes os contextos de educação e formação formais e informais, como a Educação e Formação de Adultos (EFA). No entanto, o ensino e aprendizagem da leitura para esta camada populacional, apresenta vários desafios, muitos dos quais são agravados pela escassez de materiais de leitura adequados à sua idade e experiência, o que prejudica o seu progresso e motivação (Shardlow *et al.*, 2021; Gomes, 2024).

É neste contexto que surge o projeto *iRead4Skills*, sob chancela da Faculdade de Ciências Sociais e Humanas da Universidade NOVA de Lisboa (NOVA FCSH) e coordenado por Raquel Amaro, que abrange três línguas – português, francês e castelhano. Reconhecendo a importância da leitura tanto a um nível social como individual, e ao ter como público-alvo adultos nativos com baixa literacia, os seus principais objetivos são apoiar o desenvolvimento das capacidades de leitura desse público, promovendo também novas estratégias educativas. Para isso, propõe-se a criação de um sistema inteligente de leitura, de acesso aberto e acessível na *web*, que permitirá: i) analisar a complexidade dos textos; ii) sugerir leituras de acordo com o nível de complexidade pretendido; e iii) auxiliar na simplificação de textos. Assim, foram estabelecidos três níveis de complexidade no âmbito do projeto: Muito Fácil, Fácil e Claro⁴, que irão ser discutidos no próximo capítulo.

Os trabalhos levados a cabo no contexto da bolsa deram-se no âmbito do *Work Package 3 – Complexity Classification and Data*. As tarefas desenvolvidas consistiram na compilação e anotação de um *corpus* escrito da variedade portuguesa, tendo em conta critérios e métodos de seleção, amostragem e representatividade planeados no início do projeto e que serão discutidos mais à frente. Como resultado, criou-se um *corpus* representativo dos níveis de complexidade supramencionados, englobando as necessidades e as preferências do público-alvo do projeto, compreendidas através de inquéritos deliberadamente desenhados para o efeito (Correia *et al.*, 2023; Clément *et al.*, 2023).

O *Work Package 3* contemplou ainda a elaboração de léxicos de complexidade para os níveis abrangidos no projeto (Blanco Escoda *et al.*, 2023).

⁴ Não confundir o nível Claro com a noção de linguagem clara, a qual abrange não apenas o texto, mas também o *design* e a estrutura.

Considerando o público-alvo do *iRead4Skills* – adultos com baixo nível de literacia mas que são falantes nativos da variedade portuguesa – e os níveis de complexidade tratados – que dizem respeito aos três primeiros níveis numa escala de complexidade –, é lógico e linear considerar que o vocabulário fundamental do português esteja neles incluído. Assim, serão estes léxicos de complexidade o principal recurso e objeto do trabalho proposto para a dissertação.

Por sua vez, o DLP, desenvolvido sob a chancela da ACL e coordenado por Ana Salgado, constitui um recurso lexicográfico resultante da retrodigitalização⁵ do *Dicionário da Língua Portuguesa Contemporânea* (DLPC), publicado em 2001 sob a coordenação de João Malaca Casteleiro. A versão digital do DLP, concebida para responder às exigências tecnológicas contemporâneas, foi disponibilizada *online*⁶ em abril de 2023, possibilitando o seu acesso em múltiplos dispositivos e facilitando a consulta por parte dos utilizadores.

No âmbito da bolsa, o trabalho lexicográfico em curso centra-se na atualização sistemática do DLP, envolvendo a criação e o tratamento de novas entradas lexicográficas⁷, bem como a análise, a revisão e o aperfeiçoamento de entradas já existentes, recorrendo à ferramenta lexicográfica *LeXmart*⁸ (Simões & Salgado, 2022).

Além disso, o projeto fomenta a interação com a comunidade de utilizadores, que pode sugerir novas unidades e sentidos⁹, promovendo uma atualização dinâmica e colaborativa do dicionário. Desta forma, o DLP afirma-se como uma ferramenta lexicográfica ideal para a disseminação dos resultados da investigação e para o envolvimento ativo do público.

1.3. Objetivos

No seguimento das atividades desenvolvidas nos dois projetos mencionados, e aproveitando os dados de acesso aberto deles resultantes, o objetivo principal desta

⁵ Processo de conversão de dicionários impressos para um formato digital.

⁶ <https://dicionario.acad-ciencias.pt/>

⁷ Nesta dissertação, optou-se pelo uso do termo “entrada lexicográfica”, preferindo-o a outros como “artigo lexicográfico” ou “verbetes”. Além disso, será adotada a forma reduzida “entrada” ao longo do texto. (Biderman, 2001a; Salgado, 2023).

⁸ <https://lexmart.eu/>

⁹ As sugestões são analisadas com base no contexto, fundamentando-se em *corpora*, o que resulta na inclusão ou exclusão dos vocábulos.

dissertação é a elaboração de um vocabulário fundamental do português atualizado – ou seja, a identificação das unidades lexicais¹⁰ (UL) essenciais para os diversos contextos comunicativos entre falantes nativos do português –, devidamente anotado e codificado segundo a especificação TEI Lex-0¹¹ (Tasovac & Romary *et al.*, 2018).

O DLP é um dicionário de língua geral monolíngue¹², sendo destinado ao público geral. Além disto, por se tratar de um dicionário produzido por uma Academia, assume naturalmente uma orientação prescritiva, servindo como um instrumento de referência da língua portuguesa. Nesse sentido, a articulação do trabalho desenvolvido no *iRead4Skills* com um recurso lexicográfico como o DLP facilita a conceção da ferramenta proposta, que, por ser abrangente, atual e refletir as necessidades dos falantes nativos, revela-se particularmente útil em contextos pedagógicos e sociopolíticos.

Tanto quando foi possível apurar, para a variedade portuguesa, existem duas listas de léxico fundamental: o *Português Fundamental* (Bacelar *et al.*, 1984) e o *RefLex PLE* (Pinto *et. al.*, 2024). Ambas, contudo, destinam-se ao ensino de português língua não materna e consistem numa lista de lemas¹³ que, embora contenham informação gramatical – e, no caso do *RefLex PLE*, estejam organizadas por nível de proficiência linguística –, não associam a forma a nenhum significado concreto.

Assim, a presente dissertação visa o desenvolvimento de um vocabulário fundamental diferenciador, em que, para além da atualização dos léxicos/vocabulários fundamentais existentes, propõe o seu tratamento semântico, identificando os significados em causa e dividindo-os por nível de complexidade, tendo em conta falantes nativos do português com baixa literacia.

¹⁰ Unidade básica do estudo do léxico; combina aspetos formais (forma gráfica/realização fonética) e aspetos semânticos (interpretação dada num determinado momento) e é descrita num dicionário (ISO 1951, 2007).

¹¹ <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

¹² Os dicionários de língua geral monolíngues são aqueles que pretendem registar, preservar e descrever todo o repositório linguístico de uma determinada língua, abrangendo todos os sentidos possíveis (Correia, 2009; Salgado, 2021). A cada unidade lexical pode estar associada uma etiqueta específica que fornece informações sobre o seu uso, considerando contextos geográficos, socioculturais ou domínios de especialidade.

¹³ Formas canónicas associadas às formas flexionadas, por exemplo, sou/SER ou és/SER (Mendes, 2016; Amaro & Mendes, 2024). No contexto da lexicologia e lexicografia, lema é a cabeça de uma entrada lexicográfica, sendo a representação de uma unidade lexical (ISO 1951, 2007).

Para alcançar este objetivo, estabeleceram-se metas que visam a obtenção de resultados específicos, que poderão ser reutilizados em projetos futuros com objetivos semelhantes, nomeadamente:

- I. Criação de *subcorpora* que reflitam os diferentes níveis de complexidade definidos no projeto *iRead4Skills*, garantindo a representatividade e a adequação dos textos selecionados;
- II. Identificação e definição das ULs relevantes para a construção do Vocabulário Fundamental;
- III. Integração das ULs definidas com os sentidos correspondentes no DLP, assegurando uma correspondência semântica rigorosa e contextualizada;
- IV. Criação dos metadados necessários para codificação das ULs, promovendo a interoperabilidade dos dados lexicográficos;
- V. Codificação de entradas e disponibilização no DLP.

1.4. Estrutura da dissertação

Para garantir uma organização clara e coerente, esta dissertação está estruturada em seis capítulos, subdivididos em diferentes secções e subsecções.

O segundo capítulo corresponde ao enquadramento teórico, dividido em quatro secções. Na primeira, discutimos a noção de complexidade textual e apresentamos os níveis de complexidade já referidos. A segunda secção explora a relação entre léxico e complexidade, abordando os diferentes métodos para a definição de léxicos de complexidade, distinguindo os conceitos de léxico e vocabulário e introduzindo brevemente os vocabulários fundamentais. A terceira secção é dedicada à tarefa de alinhamento de sentidos, enquanto, na quarta secção, analisamos a codificação de entradas, introduzindo a norma TEI e a especificação TEI Lex-0.

O terceiro capítulo centra-se nos recursos utilizados para o desenvolvimento do Vocabulário Fundamental, abordando a compilação do *corpus* do *iRead4Skills* e a extração dos léxicos por níveis de complexidade.

No quarto capítulo, detalhamos o processo de desenvolvimento do Vocabulário Fundamental, incluindo a identificação das ULs, o alinhamento de sentidos com o DLP e a codificação das entradas.

No quinto capítulo, apresentamos a proposta final para o Vocabulário Fundamental, incluindo a proposta de *design* gráfico, a implementação do esquema de codificação e a sua visualização.

Por fim, no capítulo seis, concluímos o nosso trabalho, destacando os principais resultados e os contributos do trabalho realizado.

2. ENQUADRAMENTO TEÓRICO

Para concretizar o objetivo desta dissertação – a elaboração de um vocabulário fundamental do português atualizado –, é necessário articular conhecimentos provenientes das áreas da linguística de *corpus*, linguística computacional, da lexicologia e da lexicografia, uma vez que a definição e compreensão desses conceitos é fundamental para delimitar o enquadramento teórico do trabalho e justificar as opções metodológicas adotadas.

Nas secções e subsecções seguintes, apresentamos os conceitos e problemáticas relevantes para o nosso trabalho.

2.1. Complexidade textual

A noção de complexidade textual pode ser definida sob várias perspetivas, variando entre autores. Por exemplo, Lorge (1948) e Kincaid *et al.* (1975) defendem abordagens baseadas em critérios quantitativos, ao passo que Harley (2013) propõe uma perspetiva centrada em critérios qualitativos.

A complexidade textual pode, por um lado, ser entendida como dependente de marcas linguísticas inerentes ao texto, mas independente das dificuldades que um determinado leitor possa enfrentar (Dahl, 2007; Cunningham & Mesmer, 2014). Nesta perspetiva, cada manifestação linguística é considerada complexa ou simples dependendo da quantidade de informação que comporta (*ibidem*). Por exemplo, ao nível da palavra¹⁴, quanto maior o número de morfemas¹⁵, mais complexa será. Assim, uma palavra no singular (e.g., *mesa*) será mais simples do que uma no plural (e.g., *mesas*). No mesmo sentido, no que respeita à sintaxe, uma frase na ordem canónica de sujeito-verbo-objeto (SVO)

– *O Pedro comeu o pão.*

será sempre mais simples do que uma frase que não segue essa norma:

¹⁴ Por palavra, referimo-nos a um “item lexical pertencente a uma determinada classe, com um significado identificável ou com uma função gramatical e com forma fonológica consistente, podendo admitir variação flexional.” (Dicionário Terminológico, s.d.).

¹⁵ Por morfema, entende-se o “constituente morfológico mínimo” (Dicionário Terminológico, s.d.), ou seja, a unidade mínima dotada de significado, seja de natureza gramatical ou lexical.

– *Comeu o Pedro o pão.*

Por outro lado, a complexidade textual pode ser compreendida como dependente não apenas das manifestações linguísticas, mas também do contexto de comunicação. Neste caso, a complexidade pode variar entre leitores, sendo subjetiva e relativa (CCSS/ELA, 2010; Kusters, 2003).

Apesar das suas contribuições, ambas as perspectivas tendem a negligenciar características específicas de determinados grupos de leitores (Heilman *et al.*, 2007; Mesmer *et al.*, 2012; Flor *et al.*, 2013), o que compromete o desempenho de diferentes sistemas de análise de complexidade textual (Collins-Thompson & Calla, 2004; Vajjal *et al.*, 2016).

Por esta razão, no projeto *iRead4Skills*, adotou-se uma abordagem combinada, que considera simultaneamente as manifestações linguísticas e as características extralinguísticas dos textos. Assim, a noção de complexidade textual é definida como dependente de manifestações linguísticas e de características textuais específicas, juntamente com conhecimento extralinguístico, como o tipo de conceitos apresentados ou o contexto em que determinado texto está inserido (Monteiro *et al.*, 2023).

2.1.1. Níveis de complexidade do *iRead4Skills*

Com o objetivo de criar uma proposta para analisar e classificar a complexidade dos textos, foi necessário estabelecer diferentes níveis de complexidade textual adequados a adultos nativos com baixa literacia.

Para a definição dos níveis, nas três línguas trabalhadas no projeto, seguiram-se os seguintes passos:

1. Compilação e análise de níveis e descritores existentes em referenciais de proficiência, como o Programa de Avaliação Internacional de Competências dos Adultos (PIAAC), Association of Language Testers in Europe (ALTE) e Quadro Europeu Comum de Referência para as Línguas (CEFR).
2. Elaboração de uma proposta inicial com três níveis de complexidade textual relevantes para o público-alvo e os objetivos do projeto.
3. Revisão e adaptação da proposta inicial após a validação por profissionais de EFA.

O resultado foi um referencial de análise de complexidade que classifica os textos em três níveis – Muito Fácil, Fácil e Claro –, adequados a adultos com baixo nível de literacia.

Cada nível é representado por um conjunto de descritores linguísticos e extralinguísticos que incidem exclusivamente na complexidade do texto, e não sobre a proficiência do leitor (Monteiro *et al.*, 2023). Esses descritores são considerados de forma hierárquica ou incremental, i.e., os descritores presentes no(s) nível(is) mais baixo(s) são também esperados nos níveis acima.

Os três níveis de complexidade determinados podem ser descritos como:

- a. **Muito Fácil** – Textos compreensíveis na sua totalidade ou quase na sua totalidade por todos os falantes nativos, incluindo pessoas com níveis muito baixos de escolaridade (i.e., que não terminaram o ensino básico ~ 6.º ano) e quase sem nenhuma experiência de leitura.
- b. **Fácil** – Textos compreensíveis na sua totalidade ou quase na sua totalidade por pessoas com baixos níveis de escolaridade (i.e., que completaram o ensino básico, mas que não foram além do 9.º ano) e com pouca experiência de leitura.
- c. **Claro** – Textos compreensíveis numa primeira leitura por pessoas que completaram o 9.º ano e com uma experiência de leitura funcional¹⁶.

Além dos níveis de complexidade considerados pertinentes para o projeto, foi definido um nível de complexidade mais elevado (+ **Complexo**), que, ao atuar como um limiar superior em relação ao último nível, apoiou o processo de definição dos níveis, bem como de compilação do *corpus* (processo detalhado no próximo capítulo). A descrição detalhada dos níveis de complexidade¹⁷ para o português e os descritores relevantes encontra-se no [Anexo I](#).

¹⁶ A leitura funcional distingue-se do conceito tradicional de leitura, ao ser inserido num contexto sociocultural, económico, político e linguístico, perspetivando a leitura como uma ferramenta necessária para a aquisição e manutenção de conhecimento, respondendo a necessidades individuais e coletivas (Bellahsène, 1973; UNESCO, 2003).

¹⁷ As descrições dos níveis de complexidade para as três línguas abrangidas no projeto estão disponíveis em Monteiro *et al.* (2023) e Monteiro *et al.* (no prelo).

2.2. Léxico e complexidade

Para além dos aspetos mencionados nas subsecções anteriores, para analisar a complexidade de um texto, é fundamental considerar o léxico utilizado (Blanco Escoda *et al.*, 2023; Shardlow *et al.*, 2021; De Hertog & Tack, 2018; Dürlich & François, 2018), uma vez que a escolha das palavras influencia diretamente o grau de dificuldade percebido pelo leitor e a sua capacidade de compreensão.

A previsão do nível de complexidade de uma UL permite identificar que palavras podem ser de difícil compreensão para uma determinada camada populacional, assumindo um papel central em tarefas de avaliação de complexidade textual e de simplificação lexical. Estas tarefas, comuns na área do processamento da língua natural (PLN), visam, principalmente, atuar em processos de simplificação textual (Shardlow, 2013a; Paetzold & Specia, 2017; Grabar & Saggion, 2022; Shardlow *et al.*, 2024). Em geral, utilizam-se sinónimos extraídos de *wordnets* ou dicionários de sinónimos (Rello *et al.*, 2015; Bott *et al.*, 2017). No entanto, a falta de anotação dos sinónimos de acordo com o seu nível de complexidade pode resultar em escolhas inadequadas para o nível pretendido ou, mesmo que sejam equivalentes, a UL poderá não o ser (Shardlow, 2013b). Por exemplo, na frase

– A saia *condiz* com a blusa.

uma tentativa de simplificação poderia substituir o verbo “condizer” por “casar”, resultando em

– A saia *casa* com a blusa.

Embora “casar” seja um verbo geralmente classificado como tendo uma menor complexidade, o significado pretendido neste contexto — “combinar, condizer” — pode não ser imediatamente compreensível. No nosso trabalho, verificámos que, apesar de “casar” estar associado ao nível Muito Fácil, o sentido correspondente a esse nível é “contrair matrimónio; unir ou unirem-se através do casamento”, o que demonstra a importância de considerar o nível de complexidade das ULs e não apenas o do lema.

Além disso, para uma avaliação precisa da complexidade lexical, é crucial compreender o perfil do utilizador final, especialmente o seu nível de literacia.

No contexto do *iRead4Skills*, a definição de um léxico categorizado foi essencial para o sistema de análise de complexidade. Contudo, os recursos lexicais disponíveis para

o português, ao serem concebidos para alunos de português língua não materna (e.g., Quadro Europeu Comum de Referência para as Línguas [QECR] ou Referencial Camões PLE), não são os adequados para o projeto, por serem orientados para um público-alvo diferente do projeto – adultos nativos que possuem um léxico passivo e ativo mais amplo.

2.2.1. Métodos para a definição de léxicos de complexidade

A definição de léxicos úteis para a análise de complexidade textual pode ser realizada por meio de diferentes metodologias: (i) baseadas em *corpus*; (ii) avaliação por especialistas; (iii) inferência automática e (iv) abordagem híbrida. Nos pontos abaixo, exploramos cada uma delas e explicamos qual foi adotada no âmbito do projeto *iRead4Skills* (Blanco Escoda *et al.* 2023).

– Abordagem baseada em *corpus*

Ao se basearem em dados produzidos em contextos naturais e autênticos de comunicação, as abordagens baseadas em *corpus* partem de hipóteses já formuladas e permitem detetar e quantificar diversas informações linguísticas (Mendes, 2016), sendo um método robusto para a compilação de léxicos de complexidade.

Os *corpora* fornecem dados estatísticos que possibilitam, por exemplo, a criação de listas de palavras com informação de frequência. Partindo da premissa de que a frequência de uma palavra impacta o seu reconhecimento (Brysbaert *et al.*, 2011; Shain, 2024) e, consequentemente, a sua compreensão (Dürlich & François, 2018; Viana & Ribeiro, 2020), a regularidade com que uma palavra aparece em *corpora* representativos de uma língua tem sido um dos fatores determinantes na elaboração de listas de palavras consideradas fundamentais (Bogaards, 2008; North *et al.*, 2023). Como afirmado em Viana & Ribeiro (2020, s.p): “Se o leitor não conhecer a esmagadora maioria das palavras que encontra num texto, a compreensão fica comprometida.”

Alguns dicionários modernos, voltados para aprendentes de línguas não maternas, incluem listas de palavras baseadas em taxas de frequência. O *Oxford Learner's Dictionary*, por exemplo, oferece duas listas distintas: a *Oxford 3000* e a *Oxford 5000*. Ambas foram elaboradas considerando a frequência lexical no *Oxford English Corpus* e num *corpus* baseado em textos utilizados em cursos voltados para aprendentes adultos de inglês.

O léxico de complexidade baseado na noção de frequência é necessariamente melhorado ao recorrer a *corpora* anotados consoante o nível de dificuldade dos textos – seja por classificação manual ou automática. É, então, possível extrair listas de palavras frequentes *vs.* raras em cada nível e calcular a distribuição de frequência das palavras nos diferentes níveis de complexidade¹⁸ (Shardlow, 2013b; Dürlich & François, 2018; Blanco Escoda *et. al.*, 2023).

– Avaliação por especialistas

Embora as metodologias baseadas em *corpus* sejam atualmente as mais comuns (Maddela & Xu, 2018), as listas de frequência apresentam algumas limitações. Por exemplo, não é possível assegurar que palavras mais frequentes são necessariamente menos complexas, dado que, como discutido anteriormente, a noção de complexidade é influenciada por diversos fatores (Jauhar & Specia, 2012; Shardlow, 2014). A validação por especialistas surge como uma potencial alternativa para superar estas limitações (Shardlow, 2014; Maddela & Xu, 2018; Blanco Escoda, 2024).

Cada falante possui uma intuição sobre quais palavras são mais comuns e menos complexas na língua, resultante da sua exposição ao *input* linguístico. Este fenómeno, conhecido como familiaridade lexical (Gernsbacher, 1984; Lewellen *et al.*, 1993), contrasta com a frequência objetiva, calculada com base em análises automáticas de *corpus*. Baseando-se nesta intuição, diversos métodos de ensino de línguas centram-se na aprendizagem de palavras, frases e expressões consideradas essenciais para o aluno (Blanco Escoda *et al.*, 2023).

Este é, contudo, um método demorado, que exige uma equipa dedicada. Do mesmo modo, as intuições de cada falante podem divergir, sendo, por vezes, difícil atingir um consenso entre especialistas. Nestes casos, o especialista valida as listas geradas previamente, avaliando, para cada unidade lexical, se o nível de complexidade atribuído é adequado, com base em critérios linguísticos e contextuais previamente definidos.

– Inferência automática

¹⁸ Este método foi já utilizado em projetos como o *CEFRLex* (François *et al.*, 2014), em que se criaram ferramentas lexicais num formato legível por máquina, que descreve a distribuição da frequência das palavras nos seis níveis de dificuldade estabelecidos na escala CEFR.

Os métodos automáticos de inferência para a criação de léxicos de complexidade utilizam regras concretas e preestabelecidas para inferir a complexidade de uma palavra ou atribuí-la a um nível de complexidade específico (François *et al.*, 2014; Blanco Escoda *et al.*, 2023).

De acordo com Gala *et al.* (2013), estes métodos baseiam-se em traços lexicais, como o número de letras, fonemas, sílabas, morfemas e a frequência de afixos em cada forma, bem como em traços psicolinguísticos, como a familiaridade lexical, idade de aquisição, frequência lexical e semelhanças ortográficas ou fonológicas.

Para uma análise mais aprofundada, estes métodos recorrem a informação previamente compilada (e.g., dados sobre a idade de aquisição ou nível de concretude dos conceitos denotados) e a ferramentas e técnicas de PLN. O traço distintivo destes métodos reside, assim, na utilização de modelos de linguagem, que permitem ponderar os diferentes fatores que influenciam a complexidade de uma palavra, baseando-se em grandes volumes de dados, aumentando a precisão dos resultados (*idem*).

– Abordagem híbrida

Quando combinados, os resultados obtidos por diferentes métodos podem ser potencializados. Tendo isto em consideração, no *iRead4Skills*, mais especificamente no caso do português, foi adotada uma abordagem híbrida para compilar os léxicos por nível de complexidade:

- I. Abordagem baseada em *corpus*: utilizada para a extração de listas de lemas¹⁹ classificados por níveis de complexidade e com informação semântica e acerca da classe de palavras;
- II. Avaliação por especialistas: utilizada para validar as listas obtidas na etapa anterior.

Desta forma, foi possível colmatar algumas limitações inerentes a cada método quando aplicados isoladamente (Pintard & François, 2020; Alfter *et al.*, 2022; Blanco

¹⁹ Formas canónicas associadas às formas flexionadas, por exemplo, sou/SER ou és/SER (Mendes, 2016; Amaro & Mendes, 2024). Um lema é também a cabeça de uma entrada, sendo a representação de uma UL (ISO 1951, 2007).

Escoda *et al.*, 2023). Por exemplo, a avaliação por especialistas tende a incidir sobretudo sobre palavras lexicais, uma vez que estas contêm carga semântica e remetem para realidades extralinguísticas, negligenciando frequentemente as palavras gramaticais, cuja função principal é estabelecer relações intralinguísticas entre os restantes constituintes da frase. Desta forma, uma abordagem também baseada em *corpus* equilibra os resultados ao incluir palavras gramaticais que, além de apresentarem uma frequência alta, pertencem a classes fechadas. Simultaneamente, a avaliação por especialistas permite identificar fenómenos regulares que podem passar despercebidos em *corpora*, devido à sua baixa frequência de ocorrência.

2.2.2. Léxico vs. vocabulário

Até ao momento, não foi feita uma distinção entre o conceito de léxico e o conceito de vocabulário. No entanto, tendo em conta o possível tratamento lexicográfico associado a estas listas, é fundamental fazer a distinção entre ambos, uma vez que o léxico diz respeito ao sistema global da língua, enquanto o vocabulário refere-se a um conjunto delimitado de unidades efetivamente atestadas num determinado contexto ou uso. Esta distinção é essencial para clarificar o escopo e os objetivos do trabalho lexicográfico a desenvolver.

Por léxico, adotamos a noção de um sistema abstrato que inclui todas as palavras de uma língua, incluindo as já atestadas, as neológicas, as que já não são utilizadas e todas aquelas que poderão ser formadas através dos processos de construção de palavras de uma língua (Biderman, 2001a; Correia & Lemos, 2005; Basílio, 2006, Dicionário Terminológico, s.d.). Por outras palavras, “O léxico é constituído [...] por formas fixas e por padrões que permitem gerar novas formas” (Rio-Torto *et al.*, 2013, p. 71).

Por outro lado, utilizamos vocabulário para nos referirmos ao “conjunto factual de todos os vocábulos atestados num determinado registo linguístico” (Correia & Lemos, 2005, p. 11), que ocorre maioritariamente no nível do discurso (Biderman, 2001b).

Neste sentido, ao tratarmos de uma lista com palavras fundamentais para um público-alvo em concreto, referimo-nos a um vocabulário, composto por uma lista finita e passível de ser incrementada.

2.2.3. Vocabulários fundamentais: métodos, objetivos e limitações

Os vocabulários fundamentais resultam de investigações teóricas e estatísticas, como a estatística lexical, que, ao utilizar métodos da linguística de *corpus*, possibilitam a compilação de listas de palavras que representam o vocabulário mais frequente de uma determinada língua. Estas listas são elaboradas com base em critérios de frequência – informação extraída a partir das ocorrências num *corpus* –, complementados por um *corpus* de disponibilidade, que identifica o vocabulário temático reconhecido como essencial para a comunicação, embora possa ter baixa probabilidade de ocorrência no *corpus* de frequência (Gougenheim *et al.*, 1954; Bacelar do Nascimento *et al.*, 1984; Lopez Morales, 1986). O objetivo destas pesquisas é identificar o vocabulário mais utilizado em situações da vida quotidiana, direcionando-o, sobretudo, para o ensino de línguas estrangeiras.

No caso do português, inspirado nos estudos realizados para o Francês Fundamental (Gougenheim *et al.*, 1954) e para o Espanhol Fundamental (Lopez Morales, 1986), foi desenvolvido o Português Fundamental (Bacelar *et al.*, 1984). Este trabalho resultou numa lista composta por 2217 lemas, acompanhados de informação gramatical, como género, número, tempo e modo verbal.

Mais recentemente, foi desenvolvido o RefLex PLE (Pinto *et. al.*, 2024), o qual, seguindo uma metodologia semelhante à já apresentada, se distingue por estar organizado por níveis de proficiência linguística baseando-se no Quadro Europeu Comum de Referência para as Línguas (QECR, 2001) e no Quadro Europeu Comum de Referência para Línguas (CEFR, 2020). O recurso apresenta quatro listas, cada uma correspondente a um nível de proficiência: **A1** (731 palavras), **A2** (1341 palavras), **B1** (2172 palavras) e **B2** (2987 palavras). Além disso, ao contrário do Português Fundamental, o RefLex PLE não indica a categoria gramatical das palavras, mas inclui algumas colocações e regências.

Ainda assim, ambos os trabalhos são orientados para o ensino do português língua não materna, não existindo, tanto quanto sabemos, trabalhos semelhantes direcionados para adultos nativos com baixa literacia. Torna-se, assim, necessária uma atualização destas listas, tendo em conta um novo público-alvo.

Adicionalmente, ao apresentarem apenas uma lista lematizada, não se associam os lemas a significados concretos. Gunza (2015, p. 37), ao referir-se ao Português

Fundamental, menciona “insuficiências”, destacando que a lista “não apresentava[m] os seus contextos de uso associados às várias situações de comunicação”. Já Cornejo (2015, p. 357) aponta a “falta de correspondencia entre el lema y la unidad léxica de aprendizaje” como uma limitação comum destes trabalhos.

No nosso entender, a determinação dos significados específicos é de extrema relevância para a definição de um vocabulário fundamental, uma vez que é expectável que diferentes ULs estejam associadas a distintos níveis de complexidade e/ou proficiência.

Desta foram, além das imprecisões já identificadas por nós e outras por diversos autores – como o reconhecimento e o tratamento de arcaísmos, empréstimos ou expressões multipalavra (MWE) (Cornejo, 2015; Ávila Muñoz, 2016; Shardlow *et al.*, 2021) –, os vocabulários fundamentais destacam-se também pela ausência da conexão com o vocabulário efetivo dos falantes, não identificando concretamente qual(is) a(s) unidade(s) lexical(is) que deverá(m) ser dominada(s).

Para ultrapassarmos estas limitações, é necessário associar cada lema à UL correspondente. Para isto, tarefas de alinhamento de sentidos revelam-se vantajosas para alcançarmos resultados eficazes.

2.3. Alinhamento e integração de conteúdo lexical

Para estabelecer esta associação, propusemo-nos recorrer ao DLP, associando cada lema ao sentido correspondente ou, inversamente, atribuindo aos sentidos relevantes a informação pertencente ao vocabulário fundamental (e ao respetivo nível de complexidade).

Nas subsecções seguintes, apresentamos a tarefa de alinhamento de sentidos, juntamente com algumas dificuldades já descritas na literatura (Ahmadi *et al.*, 2020; Salgado *et al.*, 2020), bem como de outros desafios antecipados no contexto do presente trabalho.

2.3.1. Alinhamento de sentidos

As tarefas de alinhamento de sentidos (*word sense alignment* – WSA) permitem a associação entre ULs, estabelecendo relações entre sentidos²⁰ registados em dois ou mais recursos lexicográficos distintos.

Os dicionários, enquanto recursos lexicográficos por excelência e acessíveis ao público, são objetos multifacetados, com impacto a nível individual, social, cultural e político, registando não apenas as ULs de uma língua, mas também informações gramaticais e de uso (Correia, 2009; Salgado, 2023). São, assim, ferramentas valiosas neste tipo de tarefas. A tarefa de alinhamento de sentidos entre dois ou mais dicionários requer identificar sentidos correspondentes nas mesmas entradas e, em seguida, estabelecer relações semânticas (e.g., relacionado, insuficiente, restrito, parcial²¹) entre eles (Ahmadi *et al.*, 2020; Salgado *et al.*, 2020; Caselli & Strapparava, 2014).

Embora o alinhamento de sentidos seja uma tarefa útil em várias aplicações de PLN — como a desambiguação de sentidos, a extração de informação e a anotação semântica (Ahmadi *et al.*, 2020) —, enfrenta diversos desafios. Destacam-se a falta de *corpora* semanticamente anotados, inconsistências tipológicas e informacionais entre recursos lexicais, as limitações de dados a um número restrito de línguas e dificuldades de acesso a dados disponíveis (Salgado *et al.*, 2020).

Adicionalmente, tendo em conta o trabalho que tem vindo a ser desenvolvido no DLP, antecipamos a existência de múltiplos subsentidos (diferenças muito subtis de significado, resultantes de alterações ligeiras de contexto ou do acrescento de características muito específicas dentro da mesma definição), o que dificulta a associação entre o nosso vocabulário fundamental e o conteúdo do dicionário. Por exemplo, na entrada “sair”, a definição correspondente ao sentido 1 – “abandonar o local onde se está; passar do interior de um espaço fechado para o exterior” – pode ser dividida em duas definições distintas: i) “abandonar o local onde se está” e ii) “passar do interior de um espaço fechado para o exterior”.

²⁰ Por sentido, entendemos a definição que uma UL tem no léxico mental e que a distingue de todas as outras (Chaves, 2013).

²¹ “[...] relatedness, equivalence, broadness, and narrowness [...]” (Ahmadi *et al.*, 2020).

2.3.2. Desafios e metodologias no alinhamento de sentidos

Devido aos desafios anteriormente apresentados, a maioria das tarefas de alinhamento de sentidos não pode ser totalmente automatizada, especialmente tratando-se de línguas com menos recursos lexicográficos estruturados, como o português.

Um exemplo relevante foi o trabalho conduzido pelo grupo de trabalho dedicado ao alinhamento de sentidos em recursos monolíngues do projeto ELEXIS²². Ahmadi *et al.* (2020) apresentam um conjunto de dados anotados manualmente para dicionários monolíngues, em que a anotação foi feita ao nível do sentido, relacionando os sentidos entre dois dicionários através das relações semânticas acima mencionadas. No caso português, destaca-se o trabalho realizado por Salgado *et al.* (2020), cujo objetivo passou por alinhar sentidos entre o DLPC (Academia das Ciências de Lisboa, 2001) e o *Dicionário Aberto* (DA) (Simões & Farinha, 2010).

Embora esta tarefa dependa bastante de anotações manuais, alguns autores têm explorado métodos semiautomáticos, utilizando algoritmos de aprendizagem automática (*machine learning*). Por norma, estes casos concentram-se em redes léxico-conceituais, com relações semânticas bem definidas, como as *wordnets*, ou em recursos lexicográficos computacionais com hiperligações que estabelecem relações semelhantes, como o *Wiktionary*²³. Veja-se Meyer & Gurevych (2011), que apresenta um esforço de alinhamento da *WordNet* (Fellbaum, 1998) com o *Wiktionary*, ou, numa tentativa semelhante, Navigli & Ponzetto (2012), que propõem uma metodologia semiautomática para alinhar a *WordNet* (Fellbaum, 1998) com artigos da *Wikipedia*²⁴.

Segundo Caselli & Strapparava (2014, p. 414),

Different approaches to WSA [Word Sense Alignment] have been proposed and they all share some common elements, namely: i.) the extensive use of sense descriptions of the words (e.g., WordNet glosses);

²² Um dos grandes objetivos do projeto foi o de criar “gold-standard dataset”, isto é, “a dataset of instances used for learning and fitting parameters, for aligning senses across monolingual resources”, abrangendo 15 línguas diferentes, numa tentativa de alinhar dois dicionários de cada uma (Ahmadi *et al.*, 2020).

²³ <https://www.wiktionary.org/>

²⁴ https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:P%C3%A1gina_principal

and ii.) the extension of the basic sense descriptions with additional information such as hypernyms, synonyms and domain or category labels.

A utilização de redes de relações semânticas tem demonstrado ser uma estratégia eficaz para aumentar a precisão nas tarefas de alinhamento, especialmente em recursos computacionais mais complexos.

No entanto, o alinhamento de sentidos nem sempre resulta numa correspondência exata de 1:1 – uma UL extraída de um dicionário pode corresponder a dois sentidos distintos noutro dicionário ou vice-versa.

Para uma melhor compreensão da questão em análise, apresentamos alguns dos resultados obtidos no presente trabalho:

1) Um dos sentidos definidos, nesta dissertação, para o lema “antigo²⁵”, enquanto adjetivo, foi “que é de outra época”. No DLP, o primeiro sentido associado ao lema é “que existiu em épocas muito anteriores à atual”. Já o décimo sentido define o antigo como “que é relativo à Antiguidade”. Embora se tratem de sentidos distintos – no primeiro caso, refere-se genericamente a algo pertencente ao passado; no segundo, a um período histórico específico –, para um falante comum, ambos podem ser percecionados como correspondentes a um único sentido ou significado, conforme aqui proposto.

2) Da mesma forma, alguns dos sentidos extraídos podem não ter um equivalente direto no DLP. Veja-se o verbo “perder(-se)”²⁶: uma das definições propostas neste trabalho foi “ficar confuso, desorientado”²⁷. No entanto, no DLP não foi identificado qualquer sentido idêntico ou semelhante, evidenciando assim uma lacuna.

Este tipo de discrepância evidencia a complexidade inerente ao processo de alinhamento de sentidos, sublinhando a necessidade de ajustar a metodologia de forma a refletir adequadamente tanto as variações lexicais quanto os contextos de uso específicos associados a cada UL. Concluído o alinhamento de sentidos entre as ULs extraídas no presente trabalho e as constantes no DLP, torna-se fundamental integrar e disponibilizar a informação de modo estruturado, visando promover a interoperabilidade dos dados lexicográficos e facilitar a sua reutilização em diferentes contextos digitais.

²⁵ <https://dicionario.acad-ciencias.pt/pesquisa/?word=antigo>

²⁶ <https://dicionario.acad-ciencias.pt/pesquisa/?word=perder>

²⁷ Sentido associado a frases como “Perdi-me no meio de tanta informação”.

2.3.3. Integração e disponibilização de informação lexical

Com o surgimento dos dicionários digitais, que permitem superar as limitações típicas dos dicionários impressos, como restrições de espaço ou dificuldades de atualização, surgem novas possibilidades de integração e disponibilização de informação. Desta forma, torna-se fundamental garantir uma apresentação clara e estruturada dos dados, bem como assegurar a interoperabilidade entre diferentes recursos e sistemas lexicográficos (Costa *et al.*, 2021; Simões & Salgado, 2022; Salgado, 2023). No nosso caso, essa estruturação é essencial para facilitar os diversos processos envolvidos na elaboração de um vocabulário fundamental, desde o alinhamento de sentidos e o processamento dos dados pelos lexicógrafos, até à apresentação final da informação aos utilizadores.

Para alcançar estes objetivos, é crucial a adoção de normas, padrões e formatos adequados. Assim, os dicionários são modelizados e codificados em diversos formatos, entre os quais se destacam:

- A. Formatos de uso geral: texto simples, Word (doc ou docx), PDF;
- B. Formatos de dados estruturados: *Text Encoding Initiative* (TEI), *Lexical Markup Framework* (LFM), *Resource Description Framework* (RDF).

Os formatos de dados estruturados, em particular, não só possibilitam uma organização mais eficiente e sistemática dos dados, como também permitem estabelecer conexões entre diferentes fontes de informação, promovendo a interoperabilidade e a integração de recursos lexicográficos.

2.4. Codificação

A codificação de texto é o processo de converter documentos escritos para formatos eletrónicos para que possam ser processados e analisados por sistemas digitais. Segundo Vaughan (2022), para o sucesso deste processo é necessário:

- I. Transferir os materiais selecionados para um editor de texto no computador, através da digitalização da imagem do texto para um formato de texto efetivamente legível por sistemas computacionais;

- II. Codificar ou marcar o documento utilizando etiquetas (*tags*) e elementos de marcação, o que permite definir a estrutura e o significado do conteúdo;
- III. Validar a correção do documento, verificando o texto para garantir que está corretamente codificado e que não há erros de formatação;
- IV. Apresentar o documento ao utilizador através da *web* ou de outra interface, através da formatação e de dados corretamente apresentados.

Antes do processo de codificação, o texto digitalizado (ou simples) é apenas uma cadeia contínua de caracteres sem qualquer tipo de formatação (i.e., sem itálico, negrito, sublinhado...) ou estrutura semântica e/ou estrutural clara. É o utilizador que, ao ler, reconhece diferentes parágrafos e linhas, bem como nomes próprios, títulos, datas ou qualquer outro elemento contextual. Assim, para que ferramentas digitais possam processar esta informação, é necessária a codificação do texto (*idem*).

A codificação pode ser feita através de etiquetas (<etiqueta>). Existem três tipos principais de etiquetas:

- I. Etiqueta de abertura: indica o início de uma secção (e.g., <peessoa>);
- II. Etiqueta final: indica o fim de uma secção (e.g., </peessoa>);
- III. Etiqueta vazia: representa o que não exige um fecho explícito; é útil para elementos que têm um propósito específico e não precisam de texto ou elementos internos, como uma nota vazia ou a marcação de imagens (e.g., <nota/>).

A escrita destes elementos obedece a determinadas convenções. Assim, neste trabalho, além da utilização de parênteses angulares (< >), adotar-se-á a fonte Consolas para garantir uniformidade e clareza na representação dos dados.

A codificação realiza-se, portanto, através de linguagens de marcação que possibilitam a procura, extração e processamento sistemático da informação de forma eficiente. No entanto, existem diferentes tipos de linguagens de marcação, cada uma com propósitos específicos. A seguir, destacamos dois tipos principais:

- I. Marcação de visualização: utiliza linguagens de apresentação, como a HTML, cujo objetivo é a apresentação das características de formatação e visualização

(e.g., texto entre um par de asteriscos únicos é apresentado em itálico; texto entre um par de asteriscos duplos é apresentado em negrito) (*idem*);

- II. Marcação semântica: utiliza linguagens descritivas, como a *Text Encoding Initiative* (TEI), que delimitam cadeias de texto com significado.

A codificação de diferentes tipos de documentos requer a adoção de normas e diretrizes que garantam a consistência, a organização e a interoperabilidade entre sistemas. No âmbito deste trabalho, destacam-se as normas específicas para a atividade lexicográfica, pois possibilitam a estruturação da informação lexical de maneira clara, coerente e devidamente normalizada.

2.4.1. Normas para o trabalho lexicográfico

Segundo a ISO²⁸, uma norma é:

a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose [...] [they are] a formula that describes the best way of doing something.

Nesse sentido, as normas desempenham um papel fundamental na estruturação de diferentes recursos linguísticos – como terminologias, dicionários, taxonomias e ontologias –, alinhando-os aos objetivos e às características específicas de cada um.

No que diz respeito à lexicografia, as normas, para além de fornecerem um vocabulário comum, especificam quais os métodos e procedimentos para a construção e organização dos recursos lexicográficos.

²⁸ International Organization for Standardization (<https://www.iso.org/standards.html>). A ISO é uma organização internacional não governamental com o objetivo de desenvolver normas para diferentes áreas de especialidade. Os trabalhos são realizados por comités técnicos compostos por especialistas e organizações internacionais governamentais e não governamentais.

Existem várias normas desenvolvidas por diversas organizações e grupos de trabalho, cada uma com objetivos específicos. Sem nos alongarmos sobre o tema, destacamos apenas as três normas e os seus principais objetivos:

- a. ISO 15924²⁹ (2007): formaliza a apresentação e/ou representação de entradas em dicionários impressos ou eletrônicos. Atualmente, esta norma está em processo de revisão, com início em 2022 (ISO/DIS 15924)³⁰. A revisão segue a norma ISO 24613³¹ e oferece recomendações para lidar com estruturas heterogêneas em dicionários impressos e digitais, estabelecendo conceitos e definições essenciais sobre a apresentação das entradas.
- b. TEI³²: dedicada à representação digital de textos, principalmente em contextos académicos;
- c. *OntoLex-Lemon*³³: voltada para a representação de dados lexicais na *web* semântica.

2.4.2. TEI – *Text Encoding Initiative*

A *Text Encoding Initiative* (TEI) é uma das normas *de facto* mais amplamente utilizadas em projetos na área das humanidades digitais (Burnard, 2014; Vaughan, 2022), com o objetivo de codificar múltiplos formatos de texto e permitir a sua representação digital.

A TEI baseia-se na *Extensible Markup Language* (XML)³⁴ e, enquanto metalinguagem, oferece um vocabulário – composto por elementos³⁵ e atributos³⁶ – e um esquema, i.e., uma gramática que regula a codificação dos documentos. Este sistema

²⁹ <https://www.iso.org/standard/36609.html>

³⁰ <https://www.iso.org/standard/84863.html>

³¹ <https://www.iso.org/standard/82014.html>

³² *Text Encoding Initiative* (<https://tei-c.org/>)

³³ Desenvolvida pela W3C (<https://www.w3.org/standards/>)

³⁴ A XML é uma linguagem computacional de marcação que, através de diferentes marcações e etiquetas, permite explicitar detalhadamente cada parte de um texto (TEI Consortium, 2025). No entanto, não especifica o conteúdo semântico de cada marcação.

³⁵ Definido como uma “textual unit, viewed as a structural component” (TEI Consortium, 2025).

³⁶ Definido como a “information that is in some sense descriptive of a specific element occurrence but not regarded as part of its content.” (TEI Consortium, 2025).

facilita a descrição, a estruturação e a validação dos textos e das suas componentes internas.

As diretrizes da norma, as *P5: Guidelines for Electronic Text Encoding and Interchange* (TEI Consortium, 2025), definem os métodos necessários para codificar textos que possam ser lidos por computadores e outros dispositivos, detalhando as estruturas e características conceptuais de cada documento.

A TEI permite a organização e estruturação de vários géneros e tipos de texto, com esquemas de codificação adaptáveis às necessidades específicas de cada projeto. No entanto, todos os documentos TEI devem respeitar uma estrutura mínima, composta pelos seguintes elementos (Allés Torrent, 2019):

- Declaração XML: indica a versão XML e o tipo de codificação utilizado;
- Declaração do esquema: especifica as regras que deverão ser utilizadas na codificação do texto;
- Cabeçalho *TEI header*: secção dedicada aos metadados do documento;
- Corpo do documento: representado pelo elemento <text>, que pode incluir as subdivisões <front>, <body> e <back>;
- Conjunto de elementos estruturais comuns: definidos no capítulo 3 das diretrizes, como parágrafos (<p>), listas (<list>) ou referências bibliográficas (<ref>).

O capítulo 10 das diretrizes (TEI Consortium, 2025) é dedicado, especificamente, à codificação de dicionários e apresenta soluções para a codificação de:

- o *layout* original de uma página de dicionário, i.e., a organização visual das entradas (visão tipográfica);
- as propriedades de um texto modelizadas como uma sequência de *tokens*³⁷ (visão editorial), que apresenta como o conteúdo textual é estruturado e processado;
- as estruturas lexicais subjacentes, relacionadas com o conteúdo conceptual ou linguístico de um dicionário (visão lexical), que apresentam uma perspetiva mais abstrata e direcionada para lidar com o conteúdo linguístico (Costa *et al.*, 2022; Salgado, 2023).

³⁷ Cada palavra que ocorre no *corpus* (Mendes, 2016).

As normas TEI são caracterizadas pela sua flexibilidade, permitindo múltiplas formas de codificar a mesma informação, o que pode representar um desafio para a manutenção de uma estrutura consistente entre diferentes dicionários. Assim, por razões de interoperabilidade, é necessário restringir esta flexibilidade. Por este motivo, foi desenvolvida a TEI Lex-0 (Tasovac & Romary *et al.*, 2018), uma especificação da TEI especialmente concebida para a codificação de dicionários retrodigitalizados.

A TEI Lex-0 é uma especificação da TEI que tem como objetivo uma representação mais rigorosa dos recursos lexicográficos que adotam este formato. O seu desenvolvimento é conduzido pelo grupo de trabalho denominado *Lexical Resources* da DARIAH³⁸.

Com vista a uma melhor modelização dos recursos lexicográficos, a TEI Lex-0 reduz a flexibilidade característica da TEI, limitando o número de elementos disponíveis, exigindo a utilização de determinados atributos e impondo algumas restrições sintáticas (Salgado, 2023). Ainda assim, não se propõe como um substituto do capítulo 10 das diretrizes TEI, mas sim como um “format that existing TEI dictionaries can be univocally transformed to in order to be queried, visualised, or mined in a uniform way” (Romary & Tasovac, 2018).

O DLP, um dos recursos em uso no âmbito desta dissertação, embora siga as diretrizes da TEI, encontra-se atualmente em fase de conversão para a TEI Lex-0. Na subsecção seguinte, apresentamos brevemente este processo.

2.4.3. Estratégias de codificação do DLP

O DLP segue as diretrizes da TEI, o que permite adotar esquemas de codificação textual orientados para a uniformização e estruturação da informação.

Durante o processo de retrodigitalização do DLPC, procurou-se seguir um esquema TEI para a codificação dos dados (Salgado *et al.*, 2019). No entanto, apesar da ampla flexibilidade das diretrizes TEI, não existiam esquemas específicos para a codificação de determinados componentes microestruturais do DLPC, o que levou à necessidade de ajustar a sua aplicação em alguns casos (*idem*; Salgado, 2021; Simões & Salgado, 2022).

³⁸ <https://www.dariah.eu/activities/working-groups/lexical-resources/>

Ambos os fatores dificultam a reutilização dos dados. Neste contexto, e com o objetivo de promover a interoperabilidade entre sistemas, o DLP encontra-se atualmente em processo de conversão para o formato TEI Lex-0 (*idem*; Salgado *et al.*, 2019; Simões & Salgado, 2022). A TEI Lex-0, por se configurar como uma versão mais ajustada às necessidades práticas dos recursos lexicográficos, possibilita a criação de uma estrutura mais coesa e consistente, capaz de representar adequadamente todas as estruturas do DLP (Salgado, 2021).

Com o objetivo de facilitar o trabalho lexicográfico e reduzir a complexidade da informação apresentada num esquema TEI, optou-se pela utilização de um editor de conteúdos de acesso aberto que respeita as normas mencionadas: o *LeXmart* (Simões *et al.*, 2021), em vez de um editor XML convencional.

Importa salientar que, até ao momento da redação da presente dissertação, a anotação das entradas do DLP não contemplava a indicação do nível de complexidade associado a um determinado sentido. De igual modo, as diretrizes da TEI Lex-0 não preveem qualquer proposta para este tipo de codificação. Diante deste contexto, tornou-se necessário desenvolver um esquema de anotação específico que responda às necessidades da nossa proposta de codificação.

Tendo em conta os objetivos do presente trabalho e os aspetos discutidos neste capítulo, a nossa proposta visa assim:

(i) estabelecer a ligação entre o lema presente nos léxicos por nível de complexidade do *iRead4Skills* e a(s) respetiva(s) unidade(s) lexical(is), garantindo uma correspondência precisa e clara entre forma e significado;

(ii) definir um procedimento metodológico para o alinhamento da UL com o(s) sentido(s) correspondente(s) na respetiva entrada do DLP, assegurando uma correspondência semântica e contextual rigorosa; e

(iii) projetar e implementar um modelo de codificação das ULs de acordo com as normas da TEI Lex-0, garantindo a interoperabilidade dos dados lexicográficos e a sua reutilização em diferentes contextos digitais.

No capítulo seguinte, apresentamos as bases que permitiram dar início ao nosso trabalho.

3. RECURSOS PARA O DESENVOLVIMENTO DO VOCABULÁRIO FUNDAMENTAL

Conforme apresentado na introdução, os léxicos de complexidade do *iRead4Skills* constituíram o principal objeto do nosso trabalho. Estas listas foram extraídas de um *corpus* diversificado e atualizado, composto por textos provenientes de diferentes domínios e contextos comunicativos, compilado e estruturado com base em critérios rigorosos de seleção, amostragem e representatividade definidos no início do projeto.

Para uma melhor compreensão do trabalho desenvolvido no âmbito desta dissertação, torna-se essencial caracterizar o *corpus* e os léxicos de complexidade utilizados. Assim, antes de nos centrarmos na metodologia específica do nosso projeto, apresentamos, nas subsecções seguintes, os recursos que sustentam o desenvolvimento do Vocabulário Fundamental.

3.1. *iRead4Skills Dataset 1 (v. 2.0): corpora por nível de complexidade*

O *corpus* do *iRead4Skills* é um conjunto de dados de acesso aberto, com textos³⁹ que refletem os níveis de complexidade relevantes para falantes nativos adultos da variedade portuguesa com baixa literacia.

Na compilação do *corpus*, além de serem consideradas diferentes propriedades linguísticas e extralinguísticas dos textos, também se teve em conta preferências e necessidades de leitores adultos, recolhidas através de dois inquéritos: o *Reading Skills Survey* (Correia *et al.*, 2023) e o *Overall skills and gaps survey* (Clément *et al.*, 2023), respetivamente. Desta forma, visou-se garantir uma amostra que abrange situações de comunicação relevantes, além de preferências e/ou tendências temáticas para o público-alvo do projeto.

Como mencionado inicialmente, foram também compilados textos de um nível de complexidade superior aos previstos para o projeto, o nível + Complexo, com o objetivo

³⁹ Neste trabalho, consideramos o texto enquanto objeto empírico e complexo realizado numa língua natural, cuja dimensão contextual é intrínseca ao próprio texto (Coutinho, 2003; Gonçalves, 2018).

de servirem de referência para os níveis de complexidade abaixo, baliza essencial para a análise automática de complexidade.

Sendo um *corpus* com um objetivo específico, i.e., o treino de uma ferramenta automática de análise de complexidade textual, foi necessário recolher um conjunto robusto de materiais diversos que abranjam uma ampla amostra de potenciais textos que o sistema precisará analisar.

3.1.1. Design do corpus: tipologia textual

Tendo em conta os resultados obtidos nos questionários previamente mencionados, um dos nossos objetivos foi o de compilar um *corpus* variado e atual, com textos pertencentes a vários domínios e contextos comunicativos. Para alcançar esse objetivo, foi desenvolvida uma tipologia textual com base em géneros⁴⁰ apresentados em outras tipologias de *corpora* (Biber, 1988; David, 2001) e nos referenciais consultados para estabelecer os níveis de complexidade do projeto (ALTE, 2002; OECD, 2013; Alves & Lameira, 2021; CoE, 2020).

Foram definidas 11 categorias principais que refletem o objetivo comunicativo dos textos, e para cada uma delas foram consideradas 68 subcategorias. Essas subcategorias foram escolhidas de forma a garantir a diversidade de géneros, tópicos e propriedades textuais, permitindo estabelecer uma relação equilibrada entre a dimensão social e linguística dos textos (Anexo II).

Como se pode observar na Tabela 1, esta tipologia está organizada em torno de dois aspetos principais:

- I. Objetivo de comunicação e o suporte textual⁴¹ – para considerar situações relevantes nas quais a leitura está envolvida;
- II. Tipo de documento – Além de considerar diferentes géneros de texto, também reflete os tópicos, propósitos e conteúdos usuais de um

⁴⁰ Por género, entendemos um modelo, abstracto e cognitivo, socialmente reconhecido e aceite para diferentes situações de comunicação, que permite a realização do texto mediante a função da atividade social (Bronckart, 1997; Marcushi, 2003)

⁴¹ O suporte textual é imprescindível para que o género circule em sociedade, fixando o seu objetivo comunicativo e consequente atividade social, permitindo a divulgação no espaço e no tempo do texto e do género (Marcushi, 2003).

determinado texto (por exemplo, as diferentes secções informativas presentes numa páginas *web*).

Tabela 1. Excerto da tipologia textual estabelecida no projeto *iRead4Skills*.

Propósito & Meio	Documento
comunicação institucional/profissional (c/ anonimização de dados)	carta/ <i>email</i>
	relatório
	instruções
	ata
	comunicado de imprensa
	comunicado/ <i>newsletter</i>
	página <i>web</i> (acerca/sobre/missão...)
comunicação social (jornais, revistas)	editorial
	notícia
	reportagem
	entrevista
	artigo de opinião

O objetivo era compilar dois mil textos, cada um contendo entre 250 e 500 *tokens*. Para cada documento e para cada nível de complexidade, apontou-se para um total ou mínimo de 10 textos para cada nível (10 x Muito Fácil; 10 x Fácil; 10 x Claro), projetando um *corpus* com 500 000 a 600 000 *tokens*. O número de textos compilados para o nível + Complexo não foi predeterminado.

3.1.2. Critérios de seleção e amostragem

Para o processo de compilação, foram tidos em conta critérios e métodos de seleção, amostragem e representatividade planeados no início do projeto, nomeadamente:

- I. Autenticidade e contemporaneidade dos textos, para evitar textos artificialmente simplificados.
- II. Fontes de compilação – A compilação foi realizada através de *websites* legais de livrarias *online*, editoras, páginas de blogues, *websites* políticos nacionais e *websites* de jornais ou revistas.

- III. Formato do documento – Priorizámos textos em formato digital, evitando a necessidade de utilizar reconhecimento ótico de caracteres (OCR), de forma a aumentar a eficiência e a rapidez do processo⁴².
- IV. Dimensão dos textos – Na medida do possível, a dimensão dos textos compilados para cada nível é semelhante. Ou seja, se para o nível Muito Fácil existem textos curtos, para os seguintes níveis, o número de textos com essa dimensão deve ser equivalente.
- V. Diversificação de autores e tópicos – Sempre que possível, procurámos diversificar os autores e tópicos, centrando-nos no português contemporâneo. No entanto, não foram impostas quaisquer restrições temporais específicas.

O processo de compilação e tratamento dos textos foi rigoroso e repetido em todas as subcategorias até atingirmos os resultados desejados.

Primeiramente, fizemos um levantamento de *websites* e repositórios, nos quais poderíamos aceder aos diferentes tipos de texto planeados. Procurámos, assim, excertos de textos acessíveis em livrarias *online* ou outros repositórios legais.

Após esta fase, os excertos foram mantidos em formato PDF e, posteriormente, convertidos para formato TXT⁴³. Decidimos manter ambos os formatos para garantir o seu acesso e rastreabilidade, assegurando o acesso aos originais caso o URL seja removido ou se torne inacessível.

3.1.3. Validação dos dados

Uma vez que a tarefa de classificação dos textos apresenta um alto nível de subjetividade, a classificação dos textos compilados foi validada por uma equipa de especialistas, responsáveis pela definição dos descritores de complexidade e que não estiveram envolvidos no processo de compilação.

⁴² No entanto, após termos dificuldade em encontrar textos adequados aos primeiros níveis (principalmente para o nível Muito Fácil), recorremos também a fontes impressas.

⁴³ No caso de ter sido necessária a compilação de textos impressos, estes eram, primeiro, digitalizados em formato PDF e, em seguida, convertidos para o formato TXT.

O processo de compilação e validação⁴⁴ dos dados foi realizado em simultâneo, o que permitiu uma recolha mais exata e eficiente. Este processo garantiu uma comunicação constante entre as diferentes equipas, minimizando o efeito de enviesamento pessoal e assegurando a qualidade e a diversidade dos textos compilados.

Apenas os textos validados pela segunda equipa foram incluídos no *corpus*. Caso não existisse acordo entre a equipa de compilação e a equipa de validação, repetia-se o processo de recolha até que o número de textos desejado fosse alcançado⁴⁵. Todos os textos incluídos no *corpus* alcançaram uma taxa de concordância de 100 % entre equipas.

3.1.4. Codificação dos metadados

Após o processo de validação e a transformação do texto para um formato TXT, todos os excertos foram revistos e tratados de forma a normalizar questões de formatação (e.g., quebras de linha invulgares) e eliminar informação desnecessária (e.g., URL ou códigos numéricos).

Cada documento (tanto PDF como TXT) foi codificado de acordo com os seguintes critérios:

- a. Nível de complexidade;
- b. Propósito do texto;
- c. Tipo de documento;
- d. Número do texto (tendo em conta o número de textos compilados ao momento).

Para exemplificar a aplicação destes critérios, consideremos um documento classificado como nível Muito Fácil, com propósito de comunicação pessoal e pertencente ao género diário. Caso fosse o primeiro texto compilado, seria identificado da seguinte forma: N1_Pess_diar_01.

Tendo em conta o elevado número de dados, todos os textos foram organizados numa base de dados estruturada, com metadados abrangentes e detalhados registados num arquivo *Excel* separado⁴⁶. Desta forma, foi possível organizar toda a informação textual,

⁴⁴ Para uma explicação mais detalhada do processo de validação, consulte-se Reis *et al.*

⁴⁵ Note-se que todos os descritores de complexidade de cada nível foram previamente validados por um grupo focal com formadores de centros de EFA (Monteiro *et al.*, 2023)

⁴⁶ É de notar que não sentimos a necessidade de preencher todos os campos dos metadados, já que a disponibilização de cada texto varia consoante o tipo de documento (e.g. em comunicação social, os

ao mesmo tempo que se assegurava a rastreabilidade dos dados, caso os textos originais fossem removidos da *internet*. A Tabela 2 apresenta um exemplo dos metadados codificados de um dos documentos recolhidos.

Tabela 2. Conjunto de metadados considerados para o *corpus* do *iRead4Skills*.

Metadados	Informação do documento (exemplo)
Título	A demora no momento de pagar a conta
Nome do(s) autor(es)	A Lupa de Alguém
Nome do tradutor	
Idioma	Português
Editora da obra	
Local de publicação	
Data de publicação	07/09/2011
Número de edição	
ISBN (<i>International Standard Book Number</i>)	
ISSN (<i>International Standard Serial Number</i>)	
URL (se online e disponível)	https://a-lupa-de-alguem.blogs.sapo.pt/219509.html
DOI (<i>Digital Object Identifier</i>)	
Informações sobre direitos autorais	
Data de acesso	02/05/2023
Nível de dificuldade aproximado (N1, N2, N3)	N1 (Muito Fácil)
Propósito e meio	Comunicação pessoal
Tipo de documento	Diário
Número(s) da(s) página(s) extraída(s)	

documentos terão título e autor, mas já não terá o editor. Pelo contrário, em livros de ficção, os documentos terão título, autor e, também, editor).

Número de palavras da amostra	105
ID BD_PDF / TXT	N1_Pess_diar_04

Este arquivo permitiu a monitorização de diferentes aspetos relevantes para outros trabalhos que venham a utilizar este *corpus*, como informação sobre os autores dos textos, os ISBN e os URL de onde estes foram recolhidos. Adicionalmente, em análises futuras do *corpus*, poderão ser exploradas relações entre diferentes características de complexidade linguística e outras variáveis extralinguísticas.

3.1.5. Caracterização final do *corpus*

Nesta subsecção, apresentamos os resultados, começando por destacar a discrepância entre os objetivos iniciais, que previam um número específico de textos para cada categoria, e o número real de textos compilados.

Tabela 3. Total de textos compilados por categoria.

Tipologia	Objetivo inicial (número de textos)	Número de textos compilados
comunicação pessoal	120	200
comunicação institucional/profissional	210	246
comunicação social	300	473
comunicação comercial/disseminação	210	260
livro não ficção	330	406
livro ficção	180	407
livro didático	120	420
académico	210	177
político	90	91
legal	90	85
religioso	120	168

Total	1980	2933
--------------	------	------

A segunda coluna da Tabela 3 apresenta o número de textos projetado para cada categoria. A terceira coluna mostra o número final de textos compilados. Num panorama geral, o número absoluto de textos compilados foi superior ao planejado ($2933 > 1980$). No entanto, nas categorias assinaladas a vermelho, esta tendência não se verificou.

Encontrar textos correspondentes ao nível Muito Fácil foi desafiador, especialmente para categorias específicas, uma vez que, em muitos casos, os textos não estavam disponíveis ou eram demasiado curtos. Assim, quando apenas textos curtos estavam disponíveis, foi necessário reunir excertos adicionais para algumas categorias de forma a alcançarmos o número esperado de *tokens*.

Consequentemente, tivemos de compilar mais textos para outros níveis ou categorias para equilibrar a quantidade de textos em algumas destas.

Após a seleção das fontes, compilação, processamento e organização dos textos tendo em conta o nível (Muito Fácil, Fácil, Simples e +Complexo) e categoria de texto, o *iRead4Skills Dataset 1* (v. 2.0) consiste num total de 2933 textos, com 946 131 *tokens*.

Tabela 4. Distribuição do número de textos e *tokens* por nível de complexidade.

Níveis	Muito Fácil	Fácil	Claro	+Complexo	TOTAL
Número de textos	777	666	752	738	2933
Número de tokens	132 682	163 141	301 401	348 907	946 131

A primeira linha da Tabela 4 mostra o número de arquivos TXT, enquanto a segunda linha apresenta o número de *tokens* para cada nível.

O processo de validação levou à reclassificação e reorganização de alguns excertos. Alguns textos inicialmente classificados como um determinado nível foram reclassificados como um nível diferente ou removidos do conjunto de dados. Isso levou a que algumas categorias de certos níveis (principalmente no nível +Complexo) terminassem com mais textos do que as outras. Os dados completos da distribuição de textos e *tokens* por nível, propósito e tipo de documento estão disponíveis no [Anexo III](#).

A partir deste *corpus*, foram extraídas os léxicos por nível de complexidade. Na seguinte secção, apresentamos este processo e caracterizamos estas listas.

3.2. Léxicos por níveis de complexidade

A extração dos léxicos por níveis de complexidade foi realizada através de uma metodologia híbrida.

Os léxicos de complexidade foram extraídos utilizando uma metodologia baseada em *corpus*. As listas de lemas por níveis de complexidade foram, depois, processadas pelo sistema XIP (*XEROX Incremental Parsing*), do INESC-ID Lisboa, com anotação de *part-of-speech* (POS), frequência, informação morfossintática e sintática, assim como alguns traços semânticos, tais como: humano, animal, concreto, abstrato (dividido em ação, evento ou estado), locativo, temporal e coletivo. Os lemas e respetiva informação foram validados manualmente até a uma frequência de 10 e, posteriormente, comparados com listas de léxico extraídas de um *corpus* de materiais de português língua não materna e com o Referencial Camões PLE⁴⁷.

O resultado foi a criação de um léxico com 2729 lemas (número que engloba todos os três níveis de complexidade contemplados no projeto). As POS usadas foram nomes, advérbios, preposições, verbos, adjetivos, conjunções, interjeições, pronomes, determinantes, numerais e símbolos. A sua compreensão e subsequente classificação subentende, de novo, uma noção de hierarquia entre níveis. Neste sentido, todos os lemas pertencentes a um nível inferior são considerados também nos seguintes níveis, embora não se encontrem marcados nas listas, i.e, lemas pertencentes ao nível Muito Fácil, serão já dominados nos níveis Fácil e Claro, não sendo necessário a sua presença nas listas destes níveis. Os números relativos à distribuição das POS por nível de complexidade encontram-se na Tabela 5.

Tabela 5. Distribuição de lemas das POS por níveis de complexidade.

⁴⁷ Referencial Camões PLE – Camões – Instituto da Cooperação e da Língua

POS	Muito Fácil	Fácil	Claro	TOTAL
Nomes	410	433	701	1544
Verbos	161	147	183	491
Adjetivos	69	74	180	323
Advérbios	76	30	57	163
Determinantes	2	0	0	2
Conjunções	28	9	17	54
Interjeições	4	0	0	4
Numerais	13	0	3	16
Preposições	31	9	20	60
Pronomes	41	3	1	45
Relativos	5	1	0	6
Símbolos	0	5	7	12
TOTAL	840	711	1169	2729

No próximo capítulo, descrevemos os principais passos metodológicos adotados para o tratamento das ULs selecionadas.

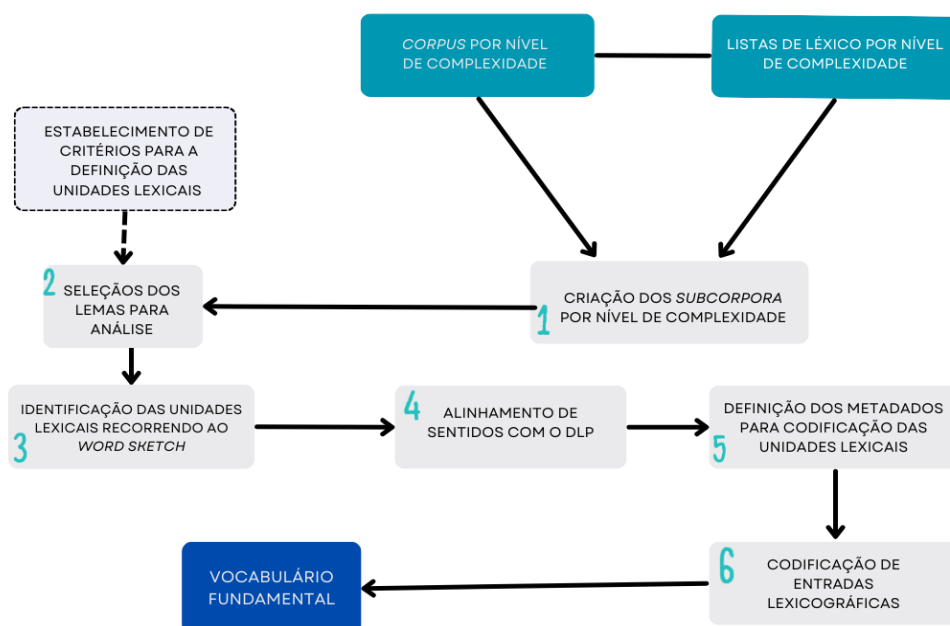
4. DESENVOLVIMENTO DO VOCABULÁRIO FUNDAMENTAL

Para o desenvolvimento de um vocabulário fundamental, com tratamento semântico e dividido por nível de complexidade, seguiram-se os seguintes passos:

1. Criação de *subcorpora*, usando o *Sketch Engine*⁴⁸, garantindo a representatividade e a adequação dos dados para análise.
2. Identificação e definição das ULs relevantes para a construção do Vocabulário Fundamental.
3. Integração das ULs definidas com os sentidos correspondentes no DLP, assegurando uma correspondência semântica rigorosa e contextualizada com os sentidos disponíveis no dicionário.
4. Criação dos metadados necessários para codificação das UL, promovendo a interoperabilidade dos dados lexicográficos.
5. Codificação de entradas.

A metodologia usada está sistematizada na Figura 1 abaixo.

Figura 1. Fluxo de trabalho para o desenvolvimento do Vocabulário Fundamental.



⁴⁸ <https://www.sketchengine.eu/>

Nas subsecções seguintes, exploraremos detalhadamente cada um destes passos, descrevendo as metodologias adotadas, os desafios enfrentados e as soluções implementadas ao longo do processo de desenvolvimento do Vocabulário Fundamental.

4.1. Criação e organização dos *subcorpora* por nível de complexidade

Para lidar com grandes volumes de textos, é fundamental recorrer a *corpus query systems* (CQS), ferramentas que possibilitam a análise sistematizada de dados linguísticos (e.g., *AntConc*⁴⁹, *Wordsmith Tools*⁵⁰). Entre essas ferramentas, destaca-se o *Sketch Engine*, uma plataforma *online* de análise de *corpus* que permite o processamento de uma grande número de dados textuais, viabilizando a compreensão do comportamento linguístico em diferentes contextos. Tendo acesso por intermédio da NOVA FCSH a esta ferramenta, e considerando os objetivos da presente análise, optou-se pela sua utilização para a criação e processamento dos *subcorpora* necessários à extração da informação relevante para o nosso trabalho.

A criação de *subcorpora* revela-se essencial para a investigação de fenómenos linguísticos específicos dentro de um *corpus* mais amplo, e visa, por exemplo, a divisão de textos com base em critérios como data, género textual ou, no nosso caso, nível de complexidade.

No contexto deste estudo, o objetivo da criação destes *subcorpora* consiste em analisar o comportamento de cada lema presente nos léxicos organizados por níveis de complexidade, de modo a identificar as ULs mais relevantes. Para a extração das ULs associadas a cada nível, foram constituídos quatro *subcorpora*, cada um correspondente aos níveis de complexidade já mencionados. A nomeação e respetiva correspondência dos *subcorpora* com os níveis de complexidade analisados são as seguintes: N1 → Muito Fácil; N2 → Fácil; N3 → Claro; N4 → +Complexo.

⁴⁹ <https://www.laurenceanthony.net/software/antconc/>

⁵⁰ <https://www.lexically.net/wordsmith/>

Para a análise dos lemas em cada *subcorpus*, seguimos uma metodologia híbrida, combinando uma extração automática das colocações⁵¹, realizada através da funcionalidade do *Word Sketch*, com uma subsequente análise manual dos resultados.

Na próxima secção, apresentamos os lemas selecionados e o processo de extração automática das colocações utilizado nesta dissertação.

4.2. Seleção dos lemas

Tendo em conta o tempo disponível para o desenvolvimento deste trabalho, a dimensão da dissertação – que obedece a critérios previamente definidos – e o facto de os léxicos de complexidade integrarem um total de 2737 lemas, tornou-se necessário restringir tanto o nível analisado como os lemas a apresentar.

Considerando a noção de hierarquia mencionada no ponto 2.1.1, optámos por centrar a nossa análise exclusivamente no nível Muito Fácil e em quatro lemas selecionados aleatoriamente a partir dos léxicos de complexidade já determinados (i.e., do *iRead4Skills*), pertencentes às principais classes abertas (nomes, verbos, adjetivos e advérbios) e fechadas (preposições e pronomes), cujo tratamento lexicográfico é também relevante. No total, iremos aqui apresentar a análise e tratamento de 24 lemas.

Ainda que a seleção tenha sido aleatória, procurámos incluir casos em que a polissemia pudesse constituir um desafio à interpretação. A Tabela 6 apresenta os lemas selecionados, juntamente com a respetiva classe gramatical.

Tabela 6. Lemas escolhidos para análise.

Classe de palavras	Lema
adjetivo	<i>antigo</i> <i>branco</i> <i>leve</i> <i>seco</i>

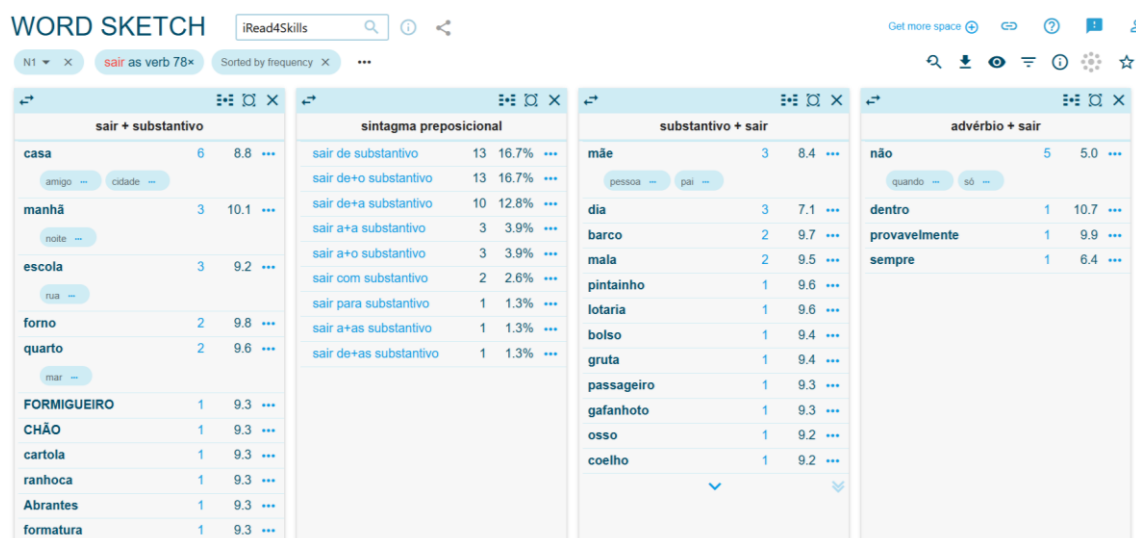
⁵¹ Por colocação referimo-nos a sequências ou combinações de duas ou mais palavras que ocorrem recorrentemente de forma simultânea e convencional (Seretan & Wehrli, 2006). Estas seguem padrões de uso consolidados, sendo condicionadas pelo idioma a que os falantes estão expostos.

advérbio	<i>agora mesmo sempre tarde</i>
nome	<i>braço conta língua pena</i>
preposição	<i>com desde durante entre</i>
pronome	<i>cada isso quanto todo</i>
verbo	<i>bater casar sair vender perder</i>

4.2.1. Identificação de ULs em *corpora*

A funcionalidade *Word Sketch* do *Sketch Engine* permite a analisar o comportamento gramatical de um determinado lema com base em dados do *corpus*, através de filtros que organizam as colocações de uma dada unidade de pesquisa, neste caso, o lema (Kilgarriff *et al.*, 2004; Sketch Engine, 2024). Os resultados desta análise são apresentados de forma estruturada, incluindo informações estatísticas relevantes, como dados de frequência e tipicidade. Desta forma, é possível visualizar desde as colocações mais frequentes e típicas até às menos recorrentes, como ilustrado na Figura 2.

Figura 2. *Word Sketch* para o lema *sair* (verbo), *subcorpus* N1.



A Figura 2 ilustra como o verbo “sair” coocorre com relevância estatística em diferentes contextos, como em posição pós-verbal com nomes (substantivos) como “casa”, “manhã” e “escola”, ou em posição pré-nominal com “mãe”, “dia”, “mala”, podendo estes nomes funcionar como sujeito ou não. Também se observa a coocorrência do verbo com sintagmas preposicionais, como “sair de”, “sair do”, “sair da”, o que permite analisar as propriedades de seleção argumental e subcategorização do verbo.

O *Word Sketch* fornece também informações estatísticas detalhadas sobre frequência e tipicidade. O primeiro número após cada colocação (por exemplo, “casa 6”) indica a frequência com que essa coocorrência aparece no *corpus*. O segundo número (e.g., “casa [...] 8.8”) refere-se à tipicidade, sendo que quanto maior a pontuação, mais típica é a colocação. Os três pontos seguintes oferecem o contexto completo da colocação, incluindo as concordâncias, e possibilitam o acesso ao *thesaurus* correspondente.

Assim, com esta funcionalidade, é possível verificar quais os contextos de ocorrência mais proeminentes nos lemas trabalhados, facilitando o tratamento das ULs correspondentes. A próxima secção aborda o processo de identificação e tratamento das ULs, recorrendo a esta funcionalidade.

4.3. Tratamento das ULs: análise e organização da informação

Os léxicos do *iRead4Skills*, embora contenham informação acerca da POS, frequência e traços semânticos, não estão semanticamente desambiguados. Assim, com o intuito de facilitar a posterior tarefa de alinhamento de sentidos, o tratamento das ULs foi orientado por alguns critérios, que passamos a enunciar:

- i) análise do tratamento lexicográfico no DLP;
- ii) proposta de definições lexicográficas⁵² (a partir de agora apenas definição) com uma estrutura semelhante às definições-modelo preexistentes no DLP⁵³;
- iii) verificação e adequação do nível de complexidade usado na definição ao nível de complexidade, para que este seja semelhante ao da UL a que a descrição se refere;
- iv) seleção de exemplos para cada UL, retirados do *subcorpus* analisado.

4.3.1. Análise do tratamento lexicográfico no DLP

No DLP, a definição consiste numa descrição semântica que contempla os vários sentidos da UL, apresentados de forma numerada e destacados a negrito, seguindo uma ordem cronológica e/ou lógica. As unidades pertencentes a classes fechadas não são objeto de descrição detalhada; em vez disso, o dicionário concentra-se na explicação do seu uso em contexto. A numeração dos diferentes sentidos de uma entrada é aplicada de forma sistemática, garantindo a organização e a clareza na apresentação da informação lexicográfica.

O DLP assinala as colocações – denominadas no próprio dicionário como combinatórias – consideradas mais relevantes. Estas são apresentadas a negrito e identificadas por um símbolo específico (🔴) no final de cada entrada ou do sentido ao qual estão associadas. As colocações são acompanhadas pela definição dos respetivos

⁵² As definições são enunciados que explicam o sentido de uma palavra ou expressão.

⁵³ Tendo em conta que “there are different lexicographic and logical ways for describing senses in a dictionary” (Ahmadi & McCrae, 2021), o que, como vimos no capítulo anterior, representa uma das dificuldades nas tarefas de alinhamento de sentidos.

significados e, sempre que necessário, são atribuídos múltiplos sentidos, devidamente distinguidos por numeração (Figura 3).

Figura 3. Colocações na entrada “língua” do DLP.

língua
lín.gua Entrada editada

nome feminino

1 língua franca
LINGÜÍSTICA, HISTÓRIA aquela que é empregue por povos que falam línguas diferentes e formada por uma mistura de elementos dessas várias línguas, a fim de os respetivos povos poderem comunicar e estabelecer relações comerciais entre si
O sabir falado até ao século XIX nos portos mediterrâneos era uma língua franca, mescla de italiano, espanhol, francês, árabe.

2 língua internacional
a que é utilizada ou foi criada para facilitar as relações internacionais, as relações entre povos que falam línguas diferentes
O inglês é hoje a língua internacional da ciência.
O latim foi durante muito tempo uma língua internacional. Para desempenhar o papel de língua internacional foi criado o esperanto.

3 dar à língua
1. INFORMAL falar, falar muito, tagarelar
Passa horas e horas a dar à língua, pouco trabalho faz.
2. dizer o que não convém que se saiba, ser indiscreto
Deu à língua e estragou o negócio ao amigo.

Adicionalmente, o DLP regista os coocorrentes privilegiados, identificados pelo símbolo . Estes são definidos como palavras que estabelecem entre si uma relação de dependência, mantendo, contudo, o seu sentido literal. O dicionário identifica os coocorrentes mais frequentes, associando-os a cada sentido específico de uma UL determinada (Figura 4).

Figura 4. Coocorrentes privilegiados associados ao sentido 1 da entrada “língua” no DLP.

nome feminino

1. **ANATOMIA** órgão carnudo, musculoso, alongado e móvel, situado na cavidade bucal que serve para a degustação, para a deglutição e, no homem e em outros animais, para a articulação de sons
 *ponta da língua; corpo da língua; freio da língua*
 *morder a língua; trincar a língua; queimar a língua; lamber com a língua*

Desta forma, para a análise dos lemas ir ao encontro do tratamento lexicográfico realizado no DLP, criámos um ficheiro *Excel* (cf. [Anexo IV](#)) com os seguintes campos: POS, ID⁵⁴, lema, sublema (combinatórias), definição, número, DLP alinhamento (de sentidos com o DLP), exemplo extraído do *Sketch Engine*, coocorrente, observações, traço(s) semântico(s).

4.3.2. Análise das ULs e elaboração de definições

Após a criação dos *subcorpora*, do estabelecimento de critérios para a definição das ULs, e da elaboração do documento de alinhamento – cujo processo será descrito na próxima secção –, estavam preenchidos todos os requisitos para o processo de análise.

Assim, no *subcorpus* N1, utilizámos a funcionalidade do *Word Sketch* para pesquisar os lemas presentes nos léxicos de complexidade e analisámos individualmente os resultados obtidos, com o objetivo de determinar e definir a(s) UL(s) em causa. Durante a primeira fase da análise, ao redigirmos as definições para cada UL, evitámos a consulta do DLP, garantindo que os resultados se baseavam exclusivamente na análise do *subcorpus*.

Na segunda fase, revimos todas as definições. Utilizando a experiência adquirida na análise e classificação dos textos durante a fase de compilação do *corpus* no projeto, juntamente com o referencial de complexidade previamente estabelecido (que inclui descrições detalhadas de cada nível – cf. Anexos I e II) , as definições foram reescritas para garantir a correspondência ao nível de complexidade atribuído à UL em questão.

Durante a identificação de combinatórias no *corpus*, considerou-se a sua recorrência em diferentes contextos comunicativos, procedendo-se igualmente à sua

⁵⁴ Cada entrada e sentido são marcados com o atributo *xml:ID* que fornece um identificador único para o elemento que contém o atributo.

análise. As combinatórias identificadas foram, assim, anotadas como sublemas e submetidas ao mesmo processo de definição aplicado aos lemas. Por exemplo, conforme ilustrado na Figura 5, no caso do lema “bater”, foi identificada a combinação “bater (as) palmas”, à qual se atribuiu a definição: “chocar as palmas das mãos uma contra a outra para demonstrar agrado, para aplaudir”.

Concomitantemente, ao identificarmos coocorrentes privilegiados, optámos por marcá-los no campo “Coocorrente”, dado que, embora não sejam expressões fixas, permitem compreender as relações de dependência mais frequentes e a grelha argumental de cada UL. Como ilustrado na Figura 6, para o lema “antigo”, identificámos os coocorrentes “Grécia Antiga”, “Lisboa antiga”, “moeda antiga”, “forma antiga” e “caminho antigo”. Estes foram devidamente identificados, marcados e associados às definições correspondentes de cada UL.

Para cada UL, retirámos um exemplo de uso do *subcorpus*, de forma a clarificar e ilustrar a definição dessa unidade.

Na Figura 5 e 6, apresentamos o resultado deste processo com os lemas “bater” e “antigo”, respetivamente .

Figura 5. Análise das ULs associadas ao lema “bater” (verbo).

POS	Lema	Sublema	Definição	Exemplo extraído do <i>Sketch Engine</i>	Coocorrente
verbo	bater		dar toques ou pancadas com alguma parte do corpo ou com um objeto	Certo dia, o pai encontrou Dinis a bater numa parede com o nariz. A tua mãe bateu à porta do quarto.	bater à porta
			chocar ou fazer chocar uma coisa em outra	A chuva bate furiosamente na janela, o vento agita as cortinas. Ele teve um acidente: bateu de frente.	
			bater (as) palmas	Ela sentia-se tão feliz que começou a bater palmas de alegria	
			chocar as palmas das mãos uma contra a outra para demonstrar agrado, para aplaudir		

Figura 6. Análise das ULs associadas ao lema “antigo” (adjetivo).

POS	Lema	Sublema	Definição	Exemplo extraído do <i>Sketch Engine</i>	Coocorrente
adjetivo	antigo		de outra época	Pégaso, que era um cavalo da Grécia antiga , voava sobre os montes à beira do mar. O anão contava-lhe histórias do passado, histórias de mouros, guerreiros, navegadores, princesas e reis antigos .	Grécia antiga Lisboa antiga
			que existe há muito tempo	Mas também nos livros antigos de Ptolomeu e nos livros dos romanos está escrito que ninguém pode passar além do Bojador.	
			que se deixou de usar	Há também antigas linhas de comboio (onde o comboio já não passa) que agora são utilizadas para andar a pé ou para praticar outros desportos.	moeda antiga; forma antiga; caminho antigo
			que já deixou de trabalhar numa certa profissão	Quem está a escrever-lhe esta carta é a sua antiga aluna do 1.º ciclo.	

Da análise dos 24 lemas analisados, identificámos 18 sublemas e 107 ULs (cf. [Anexo IV](#)).

Após concluir este processo, iniciámos a tarefa de alinhamento de sentidos, a qual será detalhada na próxima secção.

4.4. Alinhamento de sentidos: processo e dificuldades

A tarefa de alinhamento de sentidos foi realizada manualmente. No mesmo documento em que analisámos as ULs, adicionámos informações relativas ao ID da entrada (ID), ao número do sentido correspondente (N.) e à definição presente no DLP (DLP alinhamento). Estes campos possibilitam estabelecer correspondências entre os sentidos identificados na nossa análise e aqueles já registados no DLP, assegurando uma análise rigorosa e sistemática. Adicionámos também um campo para observações, no qual registámos informações sobre a possível necessidade de revisão da entrada. Esse campo permitiu documentar questões relevantes, como ambiguidades, inconsistências ou lacunas nas definições do DLP, contribuindo para um tratamento lexicográfico mais rigoroso.

Na entrada do DLP correspondente ao lema analisado, procurámos definições equivalentes. Quando encontradas, o número do sentido e a definição foram apontados no *Excel*, no espaço dedicado à UL por nós identificada.

Adicionalmente, ao longo do processo, cada UL foi anotada com os traços semânticos correspondentes, considerando o contexto e as colocações extraídas do *corpus*. Os traços semânticos adotados foram os mesmos utilizados nos léxicos por complexidade, nomeadamente: humano, animal, concreto, abstrato (subdividido em ação, evento ou estado), locativo, temporal e coletivo. Esta anotação poderá ser útil para trabalhos futuros em que seja necessário realizar tarefas de alinhamento de sentidos, uma vez que, ao se dispor da anotação semântica, será possível facilitar o processo (semi)automático, conforme explicado no capítulo dois, na subsecção 2.3.2 “Desafios e metodologias no alinhamento de sentidos”.

A Figura 7, abaixo, apresenta o resultado deste processo aplicado ao lema “seco”. No [Anexo IV](#), encontram-se os resultados deste procedimento para todos os lemas propostos nesta dissertação.

Figura 7. Alinhamento de sentidos do lema “seco” (adjetivo).

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do <i>Sketch Engine</i>	Coocorrente	Observações	Traço(s) Semântico(s)
adjetivo	DLP-seco_1-10262	seco		diz-se do alimento desidratado	5	que foi submetido a um processo de eliminação da humidade, garantindo-lhe desse modo a sua conservação	Mas ela sabia mais histórias, sabia a dos figos e contava todas as voltas que eles levavam até poderem ser chamados de figos secos .	figo seco alperce seco alho seco fruto seco		abstrato_estado
				que não é fresco	3	que perdeu a humidade natural, a frescura	Caiu no escuro, por entre uma camada de fetos secos e mmsgos ainda húmidos.			abstrato_estado
				diz-se da parte do corpo que não está molhada ou hidratada	7	diz-se de parte do corpo que não é suficientemente humedecida ou lubrificada pelas secreções	Senti a garganta seca e o coração a disparar.	garganta seca		abstrato_estado
				que é caracterizado pela falta de água ou humidade	2	que é desprovido de humidade atmosférica	Afinal ainda estamos na estação seca e a chuva vai demorar algum tempo a chegar!	estação seca		abstrato_estado
				que é brusco, áspero, que é pouco amável	16	que tem ou denota falta de sensibilidade, de temura, de amabilidade ou de afabilidade	- Cangurus, cucos e macacos - foi a resposta seca .	resposta seca		abstrato_estado

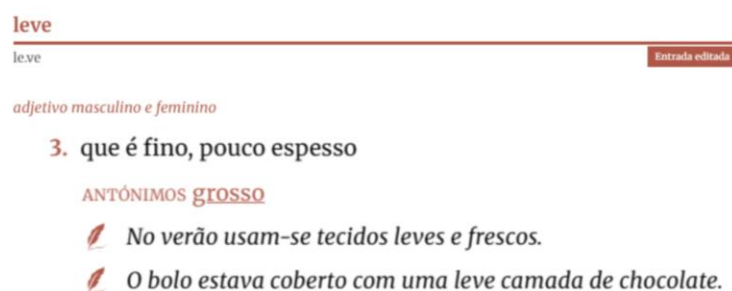
Importa observar que, como mencionado no segundo capítulo, a tarefa de alinhamento de sentidos revelou-se complexa e morosa devido a diversos fatores, entre os quais se destacam:

- I. A correspondência de uma única UL a dois sentidos no DLP (ex.: “braço” na Figura 8, colunas “Definição” e “N.” [6]);
- II. Definições não coincidentes devido a subespecificação ou sobrespecificação do sentido, ou seja, quando uma definição é mais específica do que a outra, resultando em distinções excessivamente subtis (ex.: “branco” na Figura 8);
- III. A ausência de um sentido equivalente direto no DLP para algumas ULs (ex.: “perder” na Figura 8);
- IV. A presença de subsentidos (i.e. dois sentidos distintos numa única definição) no DLP, tornando o alinhamento mais desafiante (ex.: “sair” na Figura 8, marcados a azul);
- V. Exemplos de uso no DLP que não ilustram de forma concreta e clara o sentido a que estão associados (ex.: “No verão usam-se tecidos leves e frescos” para o sentido 3 da entrada “leve” na Figura 9).

Figura 8. Dificuldades sentidas durante o processo de alinhamento.

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento
adjetivo	DLP-branco_1-86241	branco		que tem a cor da neve, do leite	1	que tem a cor da neve, da cal, do leite
nome	DLP-braco_1-795dd	braço		membro superior do corpo humano, do ombro à mão	1	cada um dos membros superiores do corpo humano, do ombro até à mão
					2	parte do membro superior compreendida entre o ombro e o cotovelo
				ir a algum lugar por diversão	4	deixar temporariamente o local onde se habita para tomar parte em atividades de caráter social ou lúdico
					21	ir para fora de casa partilhando com alguém uma diversão ou outra atividade social
	DLP-sair_1-21552	sair		deixar um lugar	1	abandonar o local onde se está; passar do interior de um espaço fechado para o exterior
					2	abandonar o local onde se está com o objetivo de ir fazer alguma coisa
					3	deixar o local onde se trabalha ou estuda, concluída a tarefa diária
					5	abandonar, um meio de transporte, a estação de embarque, rumo ao seu destino
			sair do forno	passar de fora para dentro	1	passar do interior de um espaço fechado para o exterior
				ignorar onde algo ou alguém se encontra	14	ser comprado com facilidade pelo público; ser vendável
				ficar desorientado ou distraído		não saber onde se encontra, onde pôs, deixou ou colocou
				chegar atrasado		-
				ficar sem algo que se tinha	12	não apanhar um veículo, um meio de transporte, por ter chegado atrasado
			a não perder	algo que deve ser visto ou feito		deixar de ter, de possuir, uma característica, qualidade, defeito; ficar privado de; ficar sem
			perder a conta	deixar de saber a quantidade de algo		-
						deixar de saber quantos são

Figura 9. Entrada “leve”: exemplo de uso que não ilustra claramente a definição.



Tendo em conta estas dificuldades, torna-se evidente a necessidade de uma revisão contínua dos recursos lexicográficos, nomeadamente do DLP, de modo a garantir uma estrutura mais coerente e funcional.

A complexidade e a morosidade do alinhamento demonstram que certas subdivisões, lacunas e ambiguidades no DLP comprometem a clareza e a precisão da representação lexical, exigindo, assim, uma reavaliação criteriosa. Assim, com o intuito de contribuir para a melhoria e revisão do DLP, analisámos a entrada “bater”, uma vez que esta é bem ilustrativa da maioria dos desafios anteriormente apresentados.

A entrada original, sem qualquer intervenção da nossa parte, encontra-se no [Anexo V](#). A entrada revista encontra-se já disponível no DLP⁵⁵ e pode ser consultada na íntegra no [Anexo VI](#). As principais alterações foram as seguintes:

- I. **Substituição de exemplos:** Procedeu-se à remoção de exemplos que, além de não refletirem o uso contemporâneo da língua, veiculavam mensagens socialmente problemáticas, ao sugerirem uma relação ambígua ou contraditória entre violência e afeto. Esta medida teve como objetivo garantir que o conteúdo lexicográfico reflita o uso real da língua, evitando reforçar narrativas potencialmente nocivas. Foram, assim, eliminados enunciados como: “O marido era uma pessoa muito agressiva, [...] quis-lhe bater, etc., mas enfim a coisa lá serenou” e “O pai bateu-te, mas o pai é teu amigo...”, ambos anteriormente associados ao sentido 1 da Figura 10.
- II. **Simplificação das definições:** As definições foram reformuladas com o objetivo de eliminar informações acessórias que dificultavam a compreensão do sentido central. Esta abordagem visou garantir uma leitura mais clara e objetiva. Por exemplo, o sentido 1 na Figura 10 (“dar pancadas, com as mãos ou com um objeto, um instrumento, numa pessoa ou num animal, a fim de magoar, castigar”) foi simplificado para o sentido 2 na Figura 11 (“agredir fisicamente”).

Figura 10. Entrada “bater” antes da revisão: sentidos 1 a 4.

⁵⁵ <https://dicionario.acad-ciencias.pt/pesquisa/?word=bater>

bater

Entrada editada

verbo

1. dar pancadas, com as mãos ou com um objeto, um instrumento, numa pessoa ou num animal, a fim de magoar, castigar

SINÓNIMOS agredir; dar pancada; sovar

«O marido era uma pessoa muito agressiva, [...] quis-lhe bater, etc, mas enfim a coisa lá serenou.» (PF, p. 151)

«Os irmãos [...] batiam nele como em cão malhadiço.» (Aquilino, *Terras do Demo*, p. 36)

«O pai bateu-te, mas o pai é teu amigo...» (Nemésio, *Mau Tempo*, p. 69)

2. INFORMAL obter a vitória, causando a derrota do inimigo ou do adversário

SINÓNIMOS vencer

ANTÓNIMOS perder

Os portugueses bateram os castelhanos em batalhas importantes.

3. INFORMAL ser melhor ou superior; ultrapassar em qualidade, perícia, capacidade

Este vinho bate o do ano passado em sabor e odor.

Em organização ninguém o bate.

4. ultrapassar

bater o recorde

Os números relativos ao crescimento económico bateram os níveis atingidos no ano passado.

Figura 11. Entrada “bater” depois da revisão: sentidos 1 a 4.

bater

ba.ter

Entrada nova

verbo

1. dar pancadas ou golpes em

SINÓNIMOS golpear

Ela bateu na mesa com força para chamar a atenção.

Os músicos bateram nos tambores com energia.

2. agredir fisicamente

SINÓNIMOS espancar; sovar; surrar

«Os irmãos [...] batiam nele como em cão malhadiço» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 36)

3. dar um ou mais pequenos toques ou batidas em

«Um espectador sentado na bancada de cima, [...] bateu-lhe levemente no ombro: — Pst... Oiça lá, cavalheiro!» (Vitorino Nemésio, *Mau Tempo*, p. 445)

«Bateram à minha porta, / Fui abrir, não vi ninguém.» (Manuel Bandeira, *Obras Poéticas*, 1956, p. 364)

4. golpear repetidamente algo para remover sujidade, poeira ou auxiliar na limpeza

bater os tapetes; bater as almofadas

III. **Junção de sentidos:** Sentidos com comportamento semântico semelhante foram agrupados. Esta reestruturação teve como objetivo tornar a organização da entrada

mais clara e funcional, reduzindo redundâncias e facilitando a consulta por parte do utilizador. Veja-se, por exemplo, os sentidos 3 (“ser melhor ou superior; ultrapassar em qualidade, perícia, capacidade”) e 4 (“ultrapassar”) da Figura 10, , que foram reunidos no sentido 15 (“ser melhor ou superior”) da Figura 12.

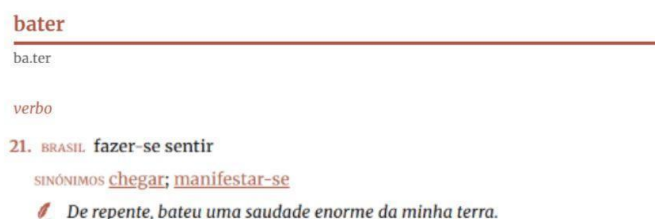
- IV. **Adição de sentidos:** Foram adicionados sentidos em falta, como o sentido 20 na Figura 12, com o objetivo de garantir uma cobertura mais completa e representativa do léxico.
- V. **Correção das etiquetas de uso**⁵⁶: As etiquetas de uso foram ajustadas de modo a refletir com maior rigor o uso efetivo dos sentidos. Por exemplo, a etiqueta de restrição geográfica associada ao sentido 21 na Figura 13 foi removida no sentido 18 da Figura 12, uma vez que esse uso está devidamente atestado em Portugal.

Figura 12. Entrada “bater” depois da revisão: sentidos 15 a 20.

-
15. ser melhor ou superior
 SINÓNIMOS superar; ultrapassar
 ☞ *bater o recorde*
 🍷 *Este vinho bate o do ano passado em sabor e odor.*
 🍷 *Em organização ninguém o bate.*
16. fabricar (moeda, medalha, etc.) imprimindo o cunho
 SINÓNIMOS cunhar
17. fechar (porta, tampa, etc.) com força
 📖 «ouvia arredar cadeiras e alguém entrar para a casa de banho e bater a porta com estrondo.»
 (M. S. Lourenço, *O Desequilibrista*, 1960, p. 109)
18. despertar determinado sentimento
 SINÓNIMOS manifestar-se
 ☞ *bater ao/no coração de alguém*
 🍷 *De repente, bateu uma saudade enorme da minha terra.*
19. fazer com que a caça se levante ou corra; fazer uma batida
 SINÓNIMOS caçar
 📖 «Homem de seiscentos diabos aquele padre Francisco, tão escoteiro nas suas obrigações de pároco como, pelos montes, a bater as lebres!» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 211)
20. DESPORTO chutar a bola num lance de bola parada
 📖 «Ronaldo bateu forte um livre lateral, Guilherme deixou a bola escapar entre as pernas e Ramsey apareceu a confirmar o golo, ainda que tenha ficado a sensação de que a bola entraria. »
 (Mais Futebol, *Liga Campeões: Ramsey «tira» golo a Ronaldo, Douglas apura Juventus*, 06/11/2019)

⁵⁶ Por etiquetas de uso referimo-nos a marcas ou indicações utilizadas para fornecer informações adicionais sobre o contexto em que uma palavra ou sentido ocorre. No DLP, a essas etiquetas abrangem diferentes categorias, incluindo (i) registo de língua (e.g. Coloquial, Popular, Poética, etc.), (ii) uso geográfico (e.g. Regionalismo, Brasileirismo, Angolanismo, etc.), (iii) domínio (e.g. Zoologia., Botânica, Linguística, etc.) (iv) frequência (e.g. Desusado ou Pouco Usado).

Figura 13. Entrada “bater” antes da revisão: sentido 21.



VI. **Reformulação da ordenação da entrada:** Procedeu-se à reorganização da entrada, de modo a posicionar no final os sentidos que apresentam restrições de uso. Esta reestruturação tem como objetivo tornar a entrada mais intuitiva e otimizar a sua consulta. (cf. [Anexo V](#) e [Anexo VI](#).)

Com esta análise abrangente da entrada “bater”, contribuímos igualmente para a revisão e o aperfeiçoamento do DLP, evidenciando a importância de uma revisão crítica contínua para a melhoria dos recursos lexicográficos.

Após o trabalho descrito nas secções anteriores, podemos avançar para a criação dos metadados necessários à codificação das ULs.

4.5. Metadados para codificação das ULs

Dado que a TEI não prevê um esquema específico para a codificação do nível de complexidade de um determinado sentido, tornou-se necessário desenvolver um modelo de anotação que permitisse a integração desta informação.

Para tal, baseámo-nos na proposta de Salgado (2023), que propôs uma codificação para hierarquizar etiquetas de domínio segundo a TEI Lex-0.

Salgado (2023) propõe a definição de taxonomias dentro do elemento `<teiHeader>`⁵⁷, possibilitando a criação de um vocabulário controlado. Este elemento permite a estruturação e identificação dos metadados do texto codificado, garantindo uma maior precisão na sua descrição. Quanto mais completa e rigorosa for a codificação dos metadados no `<teiHeader>`, maior será o controlo sobre a identidade e natureza do texto. Esta abordagem assegura a consistência na utilização das etiquetas em diferentes componentes do dicionário, assegurando uniformidade e interoperabilidade dos dados lexicográficos.

Assim, na nossa proposta, os três níveis de complexidade foram estruturados dentro do elemento `<taxonomy>`, que está inserido em `<encodingDesc>` (descrição da codificação) e agrupado em `<classDecl>` (declaração da classificação), dentro do `<teiHeader>`. No interior desse primeiro elemento, cada nível é representado por um elemento filho `<category>`, correspondendo a uma categoria distinta. O nome de cada categoria reflete o respetivo nível em inglês, enquanto o atributo `<xml:id>` atribui uma identificação única a cada nível.

A designação de cada nível é especificada dentro do elemento `<catDesc>`, que, ao incluir o atributo `<xml:lang>`, permite identificar a língua em que cada nível será apresentado, neste caso, o português (“pt”). Considerando que pretendemos mostrar ao utilizador a abreviatura de cada nível em vez da sua forma por extenso (isto é, Muito Fácil → MF; Fácil → F; Claro → C), utilizamos também o atributo `<expan>`, que contém a forma expandida de cada abreviatura.

⁵⁷ <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teiHeader.html>

Por fim, para definir cada um dos níveis, utilizámos o elemento `<gloss>`, que permite estabelecer uma relação entre palavras ou frases, facilitando a explicação ou definição de um termo. Desta forma, asseguramos que a descrição correspondente seja apresentada ao utilizador de forma clara e acessível. As definições dos níveis de complexidade apresentadas ao público serão distintas daquelas já apresentadas, com o objetivo de facilitar a sua compreensão. Os níveis de complexidade são descritos, então, da seguinte forma:

- a. **Muito Fácil** – Textos compreensíveis na sua totalidade, ou quase, por todos os falantes nativos, incluindo pessoas com níveis muito baixos de escolaridade e quase sem nenhuma experiência de leitura.
- b. **Fácil** – Textos compreensíveis na sua totalidade, ou quase, por pessoas com baixos níveis de escolaridade e com pouca experiência de leitura.
- c. **Claro** – Textos compreensíveis numa primeira leitura por pessoas com a escolaridade obrigatória e com alguma experiência de leitura.

O esquema TEI de codificação final para indicação dos níveis de complexidade no DLP é o seguinte:

```
<teiHeader>
  <encodingDesc>
    <classDecl>
      <taxonomy xml:id="level_complexity">
        <category xml:id="level_very_easy">
          <catDesc xml:lang="pt" expans="Muito fácil">MF</catDesc>
          <gloss xml:lang="pt">Palavra que surge em textos compreensíveis na sua
totalidade, ou quase, por todos os falantes nativos, incluindo pessoas
com níveis muito baixos de escolaridade e quase sem nenhuma
experiência de leitura.</gloss>
        </category>
        <category xml:id="level_easy">
          <catDesc xml:lang="pt" expans="Fácil">F</catDesc>
          <gloss xml:lang="pt">Palavra que surge em textos compreensíveis na sua
totalidade, ou quase, por pessoas com baixos níveis de escolaridade e
com pouca experiência de leitura.</gloss>
        </category>
        <category xml:id="level_plain">
          <catDesc xml:lang="pt" expans="Claro">C</catDesc>
```

```

<gloss xml:lang="pt">Palavra que surge em textos compreensíveis numa
primeira leitura por pessoas com a escolaridade obrigatória e com
alguma experiência de leitura.</gloss>
  </category>
</taxonomy>
</classDecl>
</encodingDesc>
</teiHeader>

```

Após a definição da taxonomia no <teiHeader>, é possível aplicar cada categoria ao sentido correspondente de uma determinada entrada. Para tal, recorre-se ao atributo <ana> (analysis), inserido no elemento <sense>, indicando o nível de complexidade previamente definido na secção <taxonomy> através do respetivo xml:id.

O resultado final é o seguinte:

```

<entry>
  [...]
  <sense>
    <ana="#level_very_easy"/>
  </sense>
  <sense>
    <ana="#level_easy"/>
  </sense>
  <sense>
    <ana="#level_plain"/>
  </sense>
  [...]
</entry>

```

Esta abordagem possibilita uma organização estruturada e hierárquica, garantindo a interoperabilidade dos dados lexicográficos e facilitando a sua reutilização em diversos contextos digitais.

5. INTEGRAÇÃO DO VOCABULÁRIO FUNDAMENTAL NO DLP

5.1. Proposta de *design* gráfico

Com o objetivo de disponibilizar aos utilizadores do DLP as anotações relativas aos níveis de complexidade, será implementado uma funcionalidade que permita identificar, de forma clara e intuitiva, o nível associado a cada sentido das entradas lexicais. A responsabilidade pela implementação desta funcionalidade caberá ao assessor técnico do DLP. Para apoiar este processo, recorreremos à entrada “bater” e, utilizando a ferramenta de *design* gráfico *Canva*⁵⁸, elaborámos uma proposta de apresentação para os três níveis.

O nosso objetivo foi garantir que cada nível de complexidade fosse facilmente identificado pelo utilizador e que, ao passar o cursor sobre abreviatura de cada nível, surgisse automaticamente a respetiva descrição. Esta funcionalidade visa facilitar a compreensão e a distinção entre os diferentes níveis.

Para reforçar a ligação com o *iRead4Skills*, cada nível foi representado por uma cor presente no logótipo do projeto:

– MF → verde (#00bf63)



– F → azul (#0693e3);



– C → vermelho (#f78da7).



A proposta que desenvolvemos encontra-se representada na Figura 14⁵⁹.

⁵⁸ <https://www.canva.com/>

⁵⁹ No exemplo apresentado, a ordem gradativa das anotações é meramente ilustrativa e não deve ser interpretada como uma regularidade sistemática. Trata-se de uma coincidência, não de uma característica que se verifique consistentemente noutros casos.

Figura 14. Proposta para a visualização das entradas anotadas

bater

DLPC 2001

verbo

MF

1. dar pancadas, com as mãos ou com um objeto, um instrumento, numa pessoa

Palavra que surge em textos compreensíveis na sua totalidade, ou quase, por todos os falantes nativos, incluindo pessoas com níveis muito baixos de escolaridade e quase sem nenhuma experiência de leitura.

ar, castigar

sovar

muito agressiva, [...] quis-lhe bater, etc, mas p. 151)

e como em cão malhadiço.» (Aquilino, *Terras do Demo*, p. 36)

«O pai bateu-te, mas o pai é teu amigo...» (Nemésio, *Mau Tempo*, p. 69)

F

2. INFORMAL obter a vitória, causando a derrota do inimigo ou do adversário

SINÓNIMOS vencer

ANTÓNIMOS perder

Os portugueses bateram os castelhanos em batalhas importantes.

C

3. INFORMAL ser melhor ou superior; ultrapassar em qualidade, perícia, capacidade

Este vinho bate o do ano passado em sabor e odor.

Em organização ninguém o bate.

Note-se, no entanto, que a identificação gráfica poderá ser futuramente atualizada para refletir o símbolo final de cada um dos níveis utilizado no *iRead4Skills*, que, entretanto, ainda se encontra em fase de elaboração (veja-se uma das propostas em desenvolvimento na Figura 15).

Figura 15. Proposta para a identificação gráfica do nível Muito Fácil, ainda em fase de desenvolvimento no *iRead4Skills*.



5.2. Implementação do esquema de codificação

Os conteúdos lexicais do DLP são editados recorrendo ao *LeXmart* (Simões *et al.*, 2021), um editor de conteúdos lexicográficos de acesso aberto. A opção por um editor de conteúdos lexicográficos, em detrimento de um editor XML tradicional, justifica-se pela sua interface mais intuitiva e orientada para o utilizador, que permite introduzir, editar e validar dados lexicográficos de forma estruturada, sem a necessidade de edição direta do código. Esta abordagem facilita o trabalho colaborativo, reduz a margem de erro e acelera significativamente o processo de edição.

Figura 16. Anotação do nível de complexidade no editor lexicográfico *LeXmart*.

The screenshot shows the LeXmart editor interface. At the top, there is a header with the LeXmart logo, an 'Administration' dropdown, and a search bar. Below the header, there are buttons for 'Alterar modo' and 'Guardar'. The main content area displays the entry for 'casar' with the following fields:

- entry**: xml:lang pt, xml:id DLP-casar_1-adcfd
- dac:meta**: status edited
- form**:
 - orth**: casar
 - pron**: kez'ar
 - gramGrp**: v.
- sense**: n 1, xml:id DLP-casar_1-adcfd-1
 - def**: celebrar a união legal ou religiosa de duas pessoas; realizar a cerimónia do casamento ou matrimónio
 - example**: Pediram ao padre que os casasse.
- sense**: n 2, ana #level_very_easy, xml:id DLP-casar_1-adcfd-2
 - def**: unir-se a alguém através do casamento
 - ant**: divorciar-se
 - example**: Eles casaram-se há pouco tempo.

Na Figura 16, apresenta-se a entrada “casar” no *LeXmart*, com o segundo sentido anotado como pertencente ao nível Muito Fácil, tendo por base o esquema de codificação desenvolvido na secção anterior. Já a Figura 17, abaixo, ilustra como a anotação dos sentidos é representada no editor, facilitando a visualização da informação para o lexicógrafo e garantindo uma melhor compreensão de como será apresentada ao utilizador.

Figura 17. Visualização da marcação do nível Muito Fácil no *LeXmart*.

The screenshot shows the LeXmart interface with a red header bar containing the logo and 'Administration' menu. A search bar is on the right. The main content area displays the entry for 'casar' (verb), defined as 'celebrar a união legal ou religiosa de duas pessoas; realizar a cerimónia do casamento ou matrimónio'. A red box highlights a 'Muito Fácil' label next to a text example: 'Pediram ao padre que os casasse.' Below this, a list of examples is shown, including 'Casou a filha com um médico famoso.' and '«Precisava casar aquele rapaz. Para assentar — disse Sebastião.» (Eça, *Primo Basílio*, p. 276)'. A sidebar on the right shows a list of entries with a 'Muito Fácil' label next to the 'casar' entry.

Para identificar rapidamente as entradas marcadas com um nível de complexidade, foi criado um filtro específico no editor. Este filtro, acessível na página inicial do *LeXmart*, dentro da opção *Administration > Reports > Entries with complexity analysis*, permite reunir todas as entradas já anotadas e organizá-las para uma consulta mais eficiente (Figura 18 e Figura 19), ou para posterior disponibilização independente.

Figura 18. Filtro para consultar as entradas anotadas com o nível de complexidade no *LeXmart*.

The screenshot shows the LeXmart interface with a red header bar. The 'Administration' menu is open, showing a list of options: Statistics, Labelling Lists, Reports, Export, Pages, Draft entries, Unrevised entries, New entries, Edited entries, Updated entries, Validated entries, Soon entries, and Entries with complexity analysis. The 'Entries with complexity analysis' option is selected. The main content area displays a search bar and a list of entries. The 'Pesquisa Avar' section shows a search bar and a 'Procurar' button. The 'Filtrar por' section shows a dropdown menu with 'Marcas Geográficas' selected and a 'Procurar' button. The 'Pesquisa por Wildcard (lentol!!)' section shows a search bar and a 'Procurar' button. The 'Ferramentas' section shows a 'Lista de entradas por letra' dropdown, a 'List' button, a 'List entries from word-list file' dropdown with 'Escolher ficheiro' selected, a 'List' button, a 'Nova entrada' section with a text input and a 'Criar' button, and a 'Nova página' section with a text input and a 'Criar' button. The footer shows 'Current user: leonor.reis'.

Atualmente, a lista exibe todos os sentidos, independentemente de terem sido anotados com um nível de complexidade ou não (veja-se a entrada “antigo” na Figura 19, abaixo). Futuramente, o objetivo será que sejam exibidos apenas os sentidos que tenham sido explicitamente anotados com um nível de complexidade, tornando a análise e a recuperação da informação mais precisa e direcionada. É importante destacar que esta anotação ainda não está visível no DLP, sendo apenas possível a consulta interna por parte dos membros da equipa do dicionário. A sua disponibilização será realizada quando os três níveis estiverem completamente anotados.

Figura 19. Lista das entradas anotadas com o nível de complexidade no *LeXmart*.

Entradas com anotação de complexidade	
agora¹ a.go.ra advérbio I. expressa os seguintes sentidos: MF 1. no tempo que está a decorrer; no momento presente ou nesta ocasião ☞ Que horas são agora? Agora já comia qualquer coisinha. ☞ Ela antes não me falava mas agora já me fala. ☞ Vai tu agora que eu vou depois. ☞ «Era agora — e só agora — chegada a vez de regulamentar a atividade do praticante desportivo profissional» (DAR, 16.6.1995) MF 2. na época em que vivemos SINÓNIMOS atualmente; presentemente ANTÓNIMOS antes; antigamente ☞ Agora quase toda a gente quer viver na cidade. MF 3. deste momento em diante ANTÓNIMOS antes; antigamente ☞ Agora já podes entrar. ☞ Até aqui foi fácil, mas agora é que vão ser elas. ☞ Já está tudo esclarecido, agora é esquecer os mal-entendidos. MF 4. depois disto; dadas estas circunstâncias ou em vista disto ☞ Agora vê lá se voltas a fazer o mesmo. ☞ «Para que mandou V. Ex.^a arrancar as vinhas do Ribatejo? [...] Fez V. Ex.^a um despotismo inútil e agora. Agora quem bebe por lá todo esse vinho?» (Almeida Garrett, <i>Viagens</i>, p. 55)	
antigo¹, antiga an.ti.go adjetivo MF 1. que existiu em épocas muito anteriores à atual SINÓNIMOS primitivo; remoto ☞ Os mais antigos povos da península. ☞ Não se conhecem os textos antigos citados pelo cronista. ☞ Ainda se encontram vestígios de antigos templos pagãos. MF 2. que tem muito tempo de existência ☞ móveis antigos; louças antigas SINÓNIMOS velho ANTÓNIMOS novo; recente ☞ Coleciona livros antigos. ☞ Na aldeia há uma igreja muito antiga. 3. que vive há muitos anos SINÓNIMOS idoso; velho ANTÓNIMOS jovem; novo ☞ As pessoas mais antigas lá da terra ainda se lembram disso. 4. que há muito tempo se encontra em determinada situação ou no exercício de dadas funções	

6. CONSIDERAÇÕES FINAIS E TRABALHO FUTURO

Esta dissertação teve como objetivo o desenvolvimento do Vocabulário Fundamental da ACL, assegurando a sua atualização, anotação e codificação com base nos léxicos de complexidade do projeto *iRead4Skills*.

Para alcançar este objetivo, foi essencial compreender a noção de complexidade textual e lexical, abordar a tarefa de alinhamento de sentidos e discutir as normas aplicadas ao trabalho lexicográfico, nomeadamente a TEI Lex-0. Além disso, apresentámos as bases do nosso estudo, desde a elaboração do *corpus* do *iRead4Skills* até à definição dos léxicos por nível de complexidade. Com estas etapas consolidadas, pudemos avançar para o desenvolvimento do vocabulário, detalhando todo o processo, desde a identificação das ULs até ao alinhamento de sentidos com o DLP, culminando na codificação final das entradas analisadas e em propostas para a sua apresentação ao utilizador final. Para a presente dissertação, foram analisados 24 lemas pertencentes ao nível Muito Fácil, dos quais se identificaram 18 sublemas e 107 ULs.

Com este trabalho, contribui-se para o desenvolvimento do primeiro vocabulário fundamental em formato digital do português, inovando ao introduzir um tratamento semântico dos lemas identificados e organizando-os de acordo com o seu nível de complexidade. Ao direccionar este vocabulário para falantes nativos de português com baixos níveis de literacia, diferencia-se das abordagens existentes, que tendem a apresentar listas de lemas não desambiguadas voltadas para o ensino de português como língua não materna.

As dificuldades enfrentadas ao longo do processo de alinhamento de sentidos proporcionaram também uma reflexão crítica sobre o tratamento lexicográfico das entradas no DLP. A revisão do seu conteúdo permitiu reforçar a importância de uma atualização contínua dos recursos lexicográficos, não só para garantir maior precisão e rigor descritivo, mas também para assegurar a sua adequação às necessidades reais dos falantes e à evolução do uso da língua.

No futuro, prevê-se o desenvolvimento de uma página dedicada ao Vocabulário Fundamental numa página oficial da ACL. Esta página disponibilizará uma versão otimizada do filtro previamente implementado, permitindo uma consulta mais rápida, intuitiva e interativa. Os utilizadores poderão seleccionar um nível de complexidade específico, adaptando a pesquisa às suas necessidades, o que facilitará a sua aplicação em

diversos contextos, incluindo ensino, pesquisas e investigações linguísticas e desenvolvimento de materiais educativos.

REFERÊNCIAS BIBLIOGRÁFICAS

- Academia das Ciências de Lisboa. (2023). *Dicionário da Língua Portuguesa*. <https://dicionario.acad-ciencias.pt/> (Acesso em 01/06/2025)
- Ahmadi, S. & Mcrae, J. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. [Conference Session], *Language Resources and Evaluation Conference* (LREC 2020), Marseille. <https://aclanthology.org/2020.lrec-1.395/>
- Alfter, D., Cardon, R., François, T. (2022). A Dictionary-Based Study of Word Sense Difficulty. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with READING Difficulties (READI) within the 13th Language Resources and Evaluation Conference*, 17–24. <https://aclanthology.org/2022.readi-1.3.pdf>
- ALTE (2002). *The ALTE can do project*. www.alte.org/can_do/alte_cando.pdf
- Alves, M. J., Lameira, S. (2021). *Referencial de Competências-chave de Educação e Formação de Adultos – Nível Básico*. ANQEP.
- Amaro, R. & Mendes, A. (2024). Computational Linguistics. In: Carvalho, A. & L. Oushiro (eds.) *The Oxford Handbook of the Portuguese Language*. United States of America: Oxford University Press (no prelo).
- Amaro, R. (2014). Extracting semantic relations from Portuguese corpora using lexical-syntactic patterns. In: Calzolari N., K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 26-31 May, Reykjavik, Iceland, pp. 3001-3005. Paris: ELRA. ISBN: 978-2-9517408-8-4.
- Ávila Muñoz, A. M. (2016). El léxico disponible y la enseñanza del español. Propuesta de selección léxica basada en la teoría de los conjuntos difusos. *Journal of Spanish Language Teaching* 3 (1), 31-43. <https://doi.org/10.1080/23247797.2016.1163038>
- Bacelar do Nascimento, M. F., Garcia Marques, M. L., Segura da Cruz, M.L. (1984). *Português Fundamental. Volume II, Métodos e Documentos. Tomo 1: Inquérito de Frequência*. Lisboa: INIC.
- Bacelar do Nascimento, M. F., Rivenc, P., Segura da Cruz, M. L. (1984). *Português Fundamental. Volume II, Métodos e Documentos. Tomo 2: Inquérito de Disponibilidade*. Lisboa: INIC.

- Basílio, M. (2006). *Formação e classes de palavras no português do Brasil*. São Paulo: editora contexto
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press. doi.org/10.1017/CBO9780511621024
- Biderman, M. T. C. (2001a). A ciência da lexicografia. *ALFA: Revista De Linguística*, 28(1). São Paulo. <https://periodicos.fclar.unesp.br/alfa/article/view/3676>
- Biderman, M.T.C. (2001b). Léxico e vocabulário fundamental. *ALFA: Revista de Linguística*, 40. São Paulo. <https://periodicos.fclar.unesp.br/alfa/article/view/3994>
- Blanco Escoda, X. (2024). Léxico mejorado del español para los niveles del MCER A1, A2 y B1, y nociones metalingüísticas de base. *ONOMÁZEIN | Número especial XIV Variación(es), enseñanza y traducción: investigación(es) en fraseología*, 150-167. 10.7764/onomazein.ne14.08
- Blanco Escoda, X., Amaro, R., François, T., Garcia, M. (2023). *iRead4Skills – Baselines for complexity lexicons definition (V1.0)*. Zenodo. <https://doi.org/10.5281/zenodo.10069793>
- Bogaards, P. (2008). Frequency in Learners Dictionaries. *Proceedings of the 13th EURALEX International Congress, 1231-1236*. Espanha: Barcelona. https://www.euralex.org/elx_proceedings/Euralex2008/015_Euralex_2008_Paul%20Bogaards_Frequency%20in%20Learners%20Dictionaries.pdf
- Bott, S., Rello, L., Drndarevic, B., Saggion, H. (2012). Can spanish be simpler? LexSiS: Lexical simplification for spanish. In *Coling 2012: The 24th International Conference on Computational Linguistics*. <https://aclanthology.org/C12-1023.pdf>
- Bronckart, J.P (1997). Atividades de linguagem, textos e discursos. *Por um interacionismo sócio-discursivo*. (1a edição). EDUC-PUC-SP.
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental psychology*, 58(5), 412–424. <https://doi.org/10.1027/1618-3169/a000123>
- Burnard, L. (2014). What Is the Text Encoding Initiative?. *OpenEdition Press*, <https://doi.org/10.4000/books.oep.426>.
- Caselli, T. & Strapparava, C. (2014). Automatic Domain Assignment for Word Sense Alignment. In *Proceedings of the 2014 Conference on Empirical Methods in*

- Natural Language Processing* (EMNLP), p. 414–418. Doha: Qatar. Association for Computational Linguistics. <https://aclanthology.org/D14-1046.pdf>
- CCSS/ELA – National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.
- Chaves, R. (2013). Organização do léxico. In *Grámatica do Português*, 1, 185-212. Lisboa: Fundação Calouste Gulbenkian.
- Clément, F., Hauret, L., Amaro, R. (2023). *iRead4Skills – Overall skills and gaps survey*. Zenodo. <https://doi.org/10.5281/zenodo.10181135>
- CoE – Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing: Strasbourg. www.coe.int/langcefr
- Collins-Thompson, K. & Callan, J.P. (2004). A language modeling approach to predicting reading difficulty. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 193–200. <https://aclanthology.org/N04-1025/>
- Conselho da Europa (2001). Quadro Europeu Comum de Referência para as Línguas (QECR). Edições ASA. https://www.dge.mec.pt/sites/default/files/Basico/Documentos/quadro_europeu_comum_referencia.pdf
- Cornejo, C. T. (2015) *Léxico disponible – Procesamiento y aplicación a la enseñanza de ELE*. [Tese de Doutoramento, Universidade de Salamanca]. GREDOS – Gestión del Repositorio Documental de la Universidad de Salamanca. [10.14201/gredos.128287](https://doi.org/10.14201/gredos.128287)
- Correia, M. & Lemos, L. S. P. (2005). *Inovação lexical em português*. Lisboa: Edições Colibri e Associação de Professores de Português.
- Correia, M. (2009). *Os dicionários portugueses*. Lisboa: Editorial Caminho.
- Correia, S., Amaro, R., Gauchola, R. (2023). *iRead4Skills – Reading skills survey*. Zenodo. <https://doi.org/10.5281/zenodo.10179536>
- Costa, R., Roche, C., Salgado, A. (2022). *Standards for Representing Lexicographic Data: An Overview*. Version 1.0.0. DARIAH-Campus. [Training module]. <https://campus.dariah.eu/id/REhOykBU7pPs5zOAENdah>

- Costa, R., Salgado, A., Almeida, B. (2021). SKOS as a Key Element for Linking Lexicography to Digital Humanities. In Koraljka Golub and Ying-Hsang Liu (Eds.), *Information and Knowledge Organisation in Digital Humanities: Global Perspectives*, pp. 178-204. London: Routledge. ISBN 9780367675516.
- Coutinho, A. (2003). *Texto(s) e competência textual*. Lisboa: FCT/FCG
- Cunningham, J. W., Mesmer, H. A. (2014). Quantitative measurement of text difficulty: What's the use?. *Elementary School Journal*, 115, 255-269
- Dahl, O. (2007) Definitions of complexity. *Proceedings of the Colloquium on complexity, accuracy and fluency in second language use, learning & teaching*, 41-46.
- De Hertong, D. & Tack, A. (2018). Deep Learning Architecture for Complex Word Identification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 328–334. Association for Computational Linguistics. <https://aclanthology.org/W18-0539/>
- Dicionário Terminológico (s.d). *Dicionário Terminológico para consulta em linha*. <https://dt.dge.mec.pt/> (Acesso em 01/06/2025)
- Dürlich, L. & François, T. (2018). EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). <https://aclanthology.org/L18-1140/>
- Flor, M., Klebanov, B. B., Sheehn, k. (2013). Lexical tightness and text complexity. *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, 29–38. <https://aclanthology.org/W13-1504/>
- François, T., Gala, N., Watrin, P., Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3766–3773. Islândia: Reykjavik. European Language Resources Association (ELRA).
- Gala, N., François, T., Fairon, C. (2013). Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *eLex-Electronic Lexicography*. Estónia: Tellin <https://hal.science/hal-03194427v1>
- Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113(2), 256–281. <https://doi.org/10.1037/0096-3445.113.2.256>

- Gomes, C. (2024). *Um Dedo Borrado de Tinta. Histórias de quem não pôde aprender a ler*. Fundação Francisco Manuel dos Santos.
- Gonçalves, M. (2018). Towards a text theory: (within text linguistics). In M. A. Coutinho, A. Guilherme, J. Teixeira, & B. Carvalho (Eds.), *Grammar and text: Selected papers from the 10th and 11th Fora for Linguistic Sharing* (pp. 13-37). Cambridge Scholars Publishing. <https://www.cambridgescholars.com/resources/pdfs/978-1-5275-0568-1-sample.pdf>
- Gougenheim, G., Mich, R., Rivenc, P, Sauvageot, A (1954). *L'élaboration du français élémentaire: étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier.
- Grabar, N. & Saggion, H. (2022). Evaluation of Automatic Text Simplification: Where are we now, where should we go from here. *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, 453–463. <https://aclanthology.org/2022.jeptalnrecital-taln.47/>
- Gunza, D. (2015). *Vocabulário de base no ensino primário em Angola* [Dissertação de Mestrado, Universidade Nova de Lisboa]. RUN – Repositório da Universidade Nova de Lisboa. <http://hdl.handle.net/10362/15847>
- Harley, T. A. (2013). *Psychology of Language: From Data to Theory*. Routledge – Taylor and Francis.
- Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, 460–467. <https://aclanthology.org/N07-1058/>
- ISO 15924 (2007). *Presentation/representation of entries in dictionaries — Requirements, recommendations and information*. <https://www.iso.org/standard/36609.html>
- ISO/DIS 15924. *Presentation of Lexicographic Entries in General Language Dictionaries – Fundamentals and Recommendations*. [Em desenvolvimento].
- Jauhar, S. K., Specia, L. (2012) UOW-SHEF: SimpLex – Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features. *First Joint Conference on Lexical and Computational Semantics (*SEM)*, 477–481. Canada: Montreal.
- Kilgariff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In *Proceedings EURALEX*, 105–116. https://www.euralex.org/elx_proceedings/Euralex2004/011_2004_V1_Adam%20

KILGARRIFF,%20Pavel%20RYCHLY,%20Pavel%20SMRZ,%20David%20TU
GWELL_The%20%20Sketch%20Engine.pdf

- Kincaid, J. P., Fishburne, R. P. J., Rogers, R. L., Chissom, B. S. (1975). Derivation of new readability formulas Automated Readability Index, Fog Count and Flesch Reading Ease Formula for Navy enlisted personnel. *Institute for Simulation and Training* (56). <https://stars.library.ucf.edu/istlibrary/56>
- Klaussner, C. & Zhekova, D. (2011). Lexico-Syntactic Patterns for Automatic Ontology Building. *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, 109–114. Association for Computational Linguistics. <https://aclanthology.org/R11-2017>
- Kumar, S., Jat, S., Saxena, K. & Talukdar, P. (2019). Zero-shot Word Sense Disambiguation using Sense Definition Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp 5670–5681. <https://aclanthology.org/P19-1568/>
- Kusters, W. (2003) *Linguistic complexity: the influence of social change on verbal inflection*. Netherlands: LOT.
- Lee, D. (2001). Genres, and Registers, Text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* (5), 3, 37–72. <http://llt.msu.edu/vol5num3/lee/>
- Lewellen, M. J., Goldinger, S. D., Pisoni, D. B., Greene, B. G. (1993). Lexical familiarity and processing efficiency: individual differences in naming, lexical decision, and semantic categorization. *Journal of experimental psychology. General*, 122(3), 316–330. <https://doi.org/10.1037//0096-3445.122.3.316>
- Lopez Morales, H. (1986). La enseñanza de la lengua materna. Lingüística para maestros de español. Madrid: Playor.
- Lorge, I. (1948). The Lorge and Flesch readability formulae: A correction. *School and Society* (67), 141–142.
- Maddela, M. & Xu, X. (2018). A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3749–3760, Bélgica: Bruxelas. Association for Computational Linguistics.
- Marcuschi, L. A. (2003). A Questão do Suporte dos Gêneros Textuais. *DLCV, João Pessoa*, 1 (1), 9-26. <https://periodicos.ufpb.br/ojs2/index.php/dclv/article/view/7434>

- Meissner, F.J. (2006). Le Français fondamental et les vocabulaires de base en Allemagne. *Documents pour l'histoire du français langue étrangère ou second*, 36, 119-137. <https://doi.org/10.4000/dhfles.1198>
- Mendes, A. (2016). Linguística de *Corpus* e outros usos do *corpus* em Linguística. In: Martins, Ana Maria & Ernestina Carrilho (eds.) *Manual de Linguística Portuguesa*. Manuals of Romance Linguistics vol. 16 (Edited by Günter Holtus and Fernando Sánchez Miret). Berlin/Boston: Walter de Gruyter, pp. 224-251. <http://hdl.handle.net/10451/30696>
- Mesmer, H.A., Cunningham, J., Hiebert, E. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly* (47), 3, 235–258. <https://doi.org/10.1002/rrq.019>
- Meyer, C.M. & Gurevych, I. (2011). What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, 883–892. <https://aclanthology.org/I11-1099.pdf>
- Monteiro, R., Amaro, R., Correia, S., Pintard, A., Gauchola, R., Moutinho, M., & Blanco Escoda, X. (2023). *iRead4Skills – Complexity Levels*. Zenodo. <https://doi.org/10.5281/zenodo.10459090>
- Monteiro, R., Correia, S., Amaro, Susana, Moutinho, M., Barbosa, Sílvia, Reis, M. L. (no prelo). *Níveis e descritores de complexidade textual para adultos de baixa literacia: uma ferramenta de aferição de complexidade do projeto iRead4Skills*. (manuscrito dos autores).
- Navigli, R. & Ponzetto, S. P. (2012). Multilingual WSD with Just a Few Lines of Code: the BabelNet API. In *Proceedings of the ACL 2012 System Demonstrations*, 67–72. <https://aclanthology.org/P12-3012.pdf>
- Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 105–112. <https://aclanthology.org/P06-1014/>
- North, K., Zampieri, M., Shardlow, M. (2023). Lexical Complexity Prediction: An Overview. *ACM Computing Surveys* (55), 9, 1 – 42. <https://doi.org/10.1145/3557885>
- OECD (2013). *The survey of adult skills reader's companion*. OECD Publishing.

- OECD (2019). *Skills Matter: Additional Results from the Survey of Adult Skills*. OECD Publishing, Paris, <https://doi.org/10.1787/605ec8b3-en>
- OECD (2024). Reading Literacy. <https://www.oecd.org/en/topics/sub-issues/reading-literacy.html>
- Ontology-Lexica Community Group (2013). *Specification of Requirements/Lexico-Syntactic Patterns*. https://www.w3.org/community/ontolex/wiki/Specification_of_Requirements/Lexico-Syntactic_Patterns
- Oxford (2025). *Oxford English Dictionary*. <https://www.oed.com/?tl=true>
- Oxford (2025). *Oxford Learners Dictionary*. <https://www.oxfordlearnersdictionaries.com/>
- Paetzold, G. & Specia, L. (2017). Lexical Simplification with Neural Ranking. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 34–40. Association for Computational Linguistics. <https://aclanthology.org/E17-2006/>
- Pintard, A. & François, T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. *In Proceedings of the 1st Workshop on Tools and Resources to Empower People with REading Difficulties (READI)*, 85–92. <https://aclanthology.org/2020.readi-1.13.pdf>
- Pintard, A., François, T., Nagant de Deuxchaisnes, J., Barbosa, S., Reis, M. L., Moutinho, M., Monteiro, R., Amaro, R., Correia, S., Rodríguez Rey, S., Garcia González, M., Mu, K., & Blanco Escoda, X. (2024). iRead4Skills Dataset 1: corpora by complexity level for FR, PT and SP (v2.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.10889888>
- Pinto, J., Alexandre, N., Espírito Santo, A. (2024). *REFLEX PLE – Referencial lexical para o português língua estrangeira: um recurso para professores e investigadores de PLE/L2*. Porto: U.Porto Press.
- Reis, M. L., Barbosa, S., Moutinho, M., Monteiro, R., Correia, S., Amaro, R. (2024). Intelligent Support for Low Literacy Adults: The European Portuguese iRead4Skills Corpus. *International Journal of Emerging Technologies in Learning (iJET)*, 19(08), pp. 61–81. <https://doi.org/10.3991/ijet.v19i08.52023>.
- Rello, L., Carlini, R., Baeza-Yates, R., Bigham, J.P. (2015). A plug-in to aid online reading in spanish. *In Proceedings of the 12th Web for All Conference*, pp 1–4. <https://dl.acm.org/doi/10.1145/2745555.2746661>

- Rio-Torto, G., Soares, A. R., Pereira, I., Pereira, R., Ribeiro, S. (2013). *Gramática Derivacional do Português*. Coimbra: Imprensa da Universidade de Coimbra.
- Romary, L., Tasovac, T. (2018). *TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources*. TEI Conference and Members' Meeting, Tokyo, Japan. <https://inria.hal.science/hal-02265312/document>
- Salgado, A. (2021). *Terminological Methods in Lexicography: Conceptualising, Organising and Encoding Terms in General Language Dictionaries* [Tese de Doutoramento, Universidade Nova de Lisboa]. RUN – Repositório da Universidade Nova de Lisboa. <http://hdl.handle.net/10362/137023>
- Salgado, A., Ahmadi, S., Simões, A.; Mcrae, J., Costa, R. (2020). *Challenges of Word Sense Alignment: Portuguese Language Resources*. [Conference Session], Language Resources and Evaluation Conference (LREC 2020), Marseille.
- Salgado, A, Costa, R., Tasovac, T., Simões, A. (2019). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem *et al.* (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* (pp. 417–433). Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o. <http://hdl.handle.net/11110/1772>
- Shardlow, M. (2013a). A Comparison of Techniques to Automatically Identify Complex Words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pp. 103–109. <https://aclanthology.org/P13-3015/>
- Shardlow, M. (2013b). The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 69–77. Bulgária: Sofia. Association for Computational Linguistics. <https://aclanthology.org/W13-2908/>
- Shardlow, M. (2014). Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1583–1590. Islândia: Reykjavik. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/479_Paper.pdf
- Shardlow, M., Evans, R., Henrique Paetzold, G., Zampieri, M. (2021). SemEval-2021 Task 1: Lexical Complexity Prediction. *Proceedings of the 15th International*

- Workshop on Semantic Evaluation (SemEval-2021)*, 1–16. Association for Computational Linguistics. <https://aclanthology.org/2021.semeval-1.1/>
- Shardlow, M., North, K., Zampieri, M. (2024). Multilingual Resources for Lexical Complexity Prediction: A Review. *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, 51–59. <https://aclanthology.org/2024.determin-1.5/>
- Simões, A. & Salgado, A. (2022). Smart Dictionary Editing with LeXmart. In Klosa-Kückelhaus, A., Engelberg, Stefan, Möhrs, Christine & Storjohann, Petra (eds.). *Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag*, 423-434. ISBN 978-3-937241-87-6.
- Simões, A., Salgado, A., Costa, R. (2021). LeXmart: A platform designed with lexicographical data in mind. In I. Kosem *et al.* (Eds.), *Electronic lexicography in the 21st century: Post-editing lexicography. Proceedings of the eLex 2021 conference* (pp. 529–541). Brno: Lexical Computing CZ. ISSN 2533-5626.
- Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sketch Engine (2024). *Word Sketch — collocations and word combinations*. <https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/>
- Tasovac, T, Romary, L; Banski, P., Bowers, J., Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A., Witt, A. (2018). *TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.3*. DARIAH Working Group on Lexical Resources. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>.
- TEI Consortium (2025). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>
- UNESCO (2024). Literacy. <https://www.unesco.org/en/literacy>
- Vajjala, S., Meurers, D., Eitel, A., Scheiter, K. (2016). Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 38–48. <https://aclanthology.org/W16-4105/>
- Vaughan, N. (2022). *Introdução à codificação de textos em TEI* (parte 1), traduzido por Nicolás Llano, Programming Historian em português. <https://doi.org/10.46430/phpt0026>

Viana, F. L., & Ribeiro, I. (2020). Conhecimento lexical. In *LER: Leitura, escrita, recursos*. <https://www.projetoler.pt/texto/conhecimento-lexical>

LISTA DE FIGURAS E TABELAS

LISTA DE FIGURAS

Figura 1. Fluxo de trabalho para o desenvolvimento do Vocabulário Fundamental.

Figura 2. *Word Sketch* para o lema sair (verbo), *subcorpus* N1.

Figura 3. Colocações na entrada “língua” do DLP.

Figura 4. Coocorrentes privilegiados associados ao sentido 1 da entrada “língua” no DLP.

Figura 5. Análise das ULs associadas ao lema “bater” (verbo).

Figura 6. Análise das ULs associadas ao lema “antigo” (adjetivo).

Figura 7. Alinhamento de sentidos do lema “seco” (adjetivo).

Figura 8. Dificuldades sentidas durante o processo de alinhamento.

Figura 9. Entrada “leve”: exemplo de uso que não ilustra claramente a definição.

Figura 10. Entrada “bater” antes da revisão: sentidos 1 a 4.

Figura 11. Entrada “bater” depois da revisão: sentidos 1 a 4.

Figura 12. Entrada “bater” depois da revisão: sentidos 15 a 20.

Figura 13. Entrada “bater” antes da revisão: sentido 21.

Figura 14. Proposta para a visualização das entradas anotadas.

Figura 15. Proposta para a identificação gráfica do nível Muito Fácil, ainda em fase de desenvolvimento no *iRead4Skills*.

Figura 16. Anotação do nível de complexidade no editor lexicográfico *LeXmart*.

Figura 17. Visualização da marcação do nível Muito Fácil no *LeXmart*.

Figura 18. Filtro para consultar as entradas anotadas com o nível de complexidade no *LeXmart*.

Figura 19. Lista das entradas anotadas com o nível de complexidade no *LeXmart*.

LISTA DE TABELAS

Tabela 1. Excerto da tipologia textual estabelecida no projeto *iRead4Skills*.

Tabela 2. Conjunto de metadados considerados para o *corpus* do *iRead4Skills*.

Tabela 3. Total de textos compilados por categoria.

Tabela 4. Distribuição do número de textos e *tokens* por nível de complexidade.

Tabela 5. Distribuição de lemas das POS por níveis de complexidade.

Tabela 6. Lemas escolhidos para análise.

ANEXOS

Anexo I – Descritores linguísticos e extralinguísticos por níveis de complexidade

Nível Muito Fácil
Textos curtos e simples até cerca de 50 palavras, que se refiram ao desempenho de tarefas do quotidiano ou apresentem nova informação (ex.: textos didáticos). Conceitos simples do dia a dia, com o tópico da comunicação a ser apresentado inicialmente. Contextos básicos de comunicação: quotidianos (horários de transportes/listas/menus/informação de preço), comunicações pessoais (cartas, mensagens). Palavras simples e frequentes, comumente usadas em contextos do dia a dia (transporte, comida, família, trabalho). Uso predominante de palavras concretas que fazem referência a objetos concretos no mundo real (maçã, cadeira, fogão). Ausência de abreviaturas, siglas e estrangeirismos, à exceção daqueles muito frequentes (internet, hamburger). Palavras compostas comuns (couve-flor, guarda-chuva). Palavras com afixos frequentes (infeliz, calmamente, trabalhador). Frases curtas e simples, utilizando conjunções simples e na ordem direta (Sujeito-Verbo-Complemento). Estruturas de coordenação frequentes (copulativas, disjuntivas e adversativas) estão presentes (ex.: O Rui e/ou a sua irmã foram à festa, <u>mas</u> o João não foi). Estruturas de subordinação frequentes estão presentes (ex.: oração subordinada adverbial temporal – Desde que voltou, o Rui não dormiu). Uso predominante da voz ativa. Tempos verbais no modo Infinitivo, Indicativo e Imperativo, nomeadamente o Presente e Pretéritos Perfeito e Imperfeito. Ocorrência de construções perifrásticas comuns como <u>estar/começar/andar + a + infinitivo</u> ou <u>deixar/acabar + de + infinitivo</u>. Não figuram tempos compostos, à exceção do Futuro que é transmitido através de <u>ir + infinitivo</u>. Cadeia referencial completa sem omissão de elementos, com repetição em vez da utilização de pronomes (ex.: O Rui gosta da <u>Maria</u>, porque a <u>Maria</u> é simpática). Localização temporal linear, dada por advérbios e conjunções simples (ex.: <u>Ontem</u>, o Rui foi à praia. Os bombeiros chegaram <u>depois</u> da polícia). Ausência de figuras de estilo.

Nível Fácil
Textos curtos até 150 palavras, que são interessantes para o leitor se informar/aprender algo novo ou para momentos de lazer. Conceitos mais específicos, mas do interesse do leitor (profissional, desporto, hobbies). Contextos de comunicação profissionais e dos media (ex.: instrução profissional, notícias, anúncios, desporto). Palavras e expressões frequentes que fazem referência a assuntos do dia a dia, mas também a outros âmbitos mais específicos (ex.: curso, ano escolar). O texto contém algumas palavras que referem ideias abstratas mais frequentes (felicidade, imaginação, carinho). O texto utiliza abreviaturas comuns, depois de as ter explicado (ex.: litro – l/metro – m). O texto apenas utiliza siglas e acrónimos conhecidos (CP, IVA, RTP, SIC, TVI). Uso de estrangeirismos menos frequentes (site, souvenir). O texto utiliza palavras compostas comuns com o seu sentido a poder ser compreendido pela combinação das duas partes (óculos de sol, mal-educado, meia-idade, maus-tratos). Palavras com afixos menos frequentes (<u>pré-cozinhado</u>, <u>embalagem</u>, <u>finíssimo</u>).

Frases simples e complexas através de conjunções coordenadas e subordinadas (ex.: Os músicos estavam tocando, mas poucos clientes ouviam. Quando pararam de tocar, apenas um senhor se manifestou). Uso de orações relativas de sujeito, mas não de objeto (ex.: O senhor que vive na casa ao lado ouviu o barulho). O texto pode conter frases apresentadas numa ordem não linear (Complemento-Verbo-Sujeito) (ex.: À escola, chegaram os autocarros da visita de estudo).

Possível uso da voz passiva.

Uso de modos e tempos verbais variados.

Uso de tempos verbais compostos e de expressões verbais com ter/poder/dever (posso/devo/tenho de gostar de bolo). Modos Conjuntivo, Condicional e Gerúndio.

Expressões com os verbos permanecer, querer, poder, conseguir em tempos simples (ex.: permaneça calmo, queria almoçar, consigo ir).

Utilização do Futuro simples (ex.: ele gostará de te ver).

O texto tem alguns elementos omitidos e utiliza pronomes (ex.: Eu gostava da Maria vs. Gostava da Maria/Mas ela não gosta dele).

Sequência ou localizações temporais lineares, dadas através de advérbios de tempo ou conjunções, **mas também por tempos verbais** (ex.: Ele está a caminho e a festa já começou).

Figuras de estilo frequentes são utilizadas.

O texto faz uso da **metonímia** (ex.: beber um copo), assim como **metáforas familiares** (ex.: O meu pai tem olhos de águia).

Nível Claro

Textos com cerca de 250 palavras sobre vários tópicos, sejam estes de informação ou lazer, com os leitores a conseguirem sair fora da sua zona de conforto.

Conceitos variados que necessitam de ser interligados para a sua compreensão total (ciência, política).

Contextos comunicativos variados: lazer (literatura, reportagens); **artes** (poesia); **contextos especializados** (fóruns online, comunicação profissional formal).

Palavras e expressões menos frequentes que fazem referência a âmbitos específicos (ex.: doença cardíaca, artéria, tratamento). As palavras remetem para **conceitos mais abstratos** (envelhecimento, razão, fidelidade).

Utilização de siglas e acrónimos não conhecidos pelo leitor juntamente com o nome a que se referem (ex.: Organização das Nações Unidas (ONU), Instituto Público (IP)).

O texto utiliza abreviaturas comuns, sem explicação prévia (ex.: m, km, g., l, Exmo.).

Uso frequente de estrangeirismos comuns (ex.: link, hobby, marketing, mail, gay).

O texto contém palavras que são ambíguas e/ou pouco frequentes. (ex.: interno, representar, forma, paciente).

Palavras com sufixos e prefixos variados (ex.: nervosismo, lavável, curiosidade, altitude, superpoder, preconceituoso).

Frases frequentemente combinadas através de conjunções coordenativas e subordinativas e, por vezes, com conjunções menos comuns (ex.: apesar da confusão provocada pelos animais à solta, a situação foi controlada sem os animais serem atropelados.)

Presença frequente de ordem não linear de constituintes(sujeito-verbo-complementos) (ex.: de ar puro, todos precisamos.).

Utilização habitual da voz passiva (ex.: o almoço foi pago pela empresa.), incluindo construções impessoais comuns e passivas de -se (ex.: O trabalho faz-se bem.).

Utilização de todos os modos e tempos verbais compostos e simples, à exceção do Pretérito Mais-perfeito.

Elementos omitidos (ex.: [] mostrámos em estudos anteriores que a gordura solidifica quando [] arrefece. vs. Nós mostrámos em estudos anteriores que a gordura solidifica quando ela arrefece). Utilização de **pronomes e anáforas lexicais** para indicar o que é referido (ex.: no livro, nós damos uma definição

complete de uma tempestade de neve e como esta se forma. Também indicamos onde o fenómeno pode ser visto.)

Sequências ou localizações temporais não lineares (ex.: após analisarem as respostas dos doentes que tinham sofrido um ataque cardíaco, os médicos confirmaram que a gordura é o alimento mais perigoso.)

Uso de **figuras de estilo** menos frequentes.

Metáforas incomuns (ex.: A sociedade está doente; ele é um lobo em pele de cordeiro).

Anexo II – Tipologia textual completa

<i>Propósito e meio</i>	<i>Documento</i>	<i>Observações</i>
comunicação pessoal	bilhete	*considerar apenas exemplos presentes em literatura, devido a questões de proteção de dados. De acordo com o PLN2027, estes textos são os menos relevantes.
	E-mail/carta pessoal	
	SMS/ mensagens instantâneas (WhatsApp...)	
	lista/agenda (de compras, de tarefas...)	
	diário	
comunicação institucional/profissional	carta/e-mail	Variar o tipo de instituição e a área de trabalho. **Instruções dadas aos colaboradores (como aceder ao e-mail profissional...).
	relatório	
	instruções**	
	ata	
	comunicado de imprensa	
	comunicado/ <i>newsletter</i>	
	página <i>web</i> (acerca de nós/missão...)	
comunicação social (jornais, revistas)	editorial	variar tópico, datas...
	notícia	
	reportagem	
	entrevista	
	artigo de opinião	
	artigo de divulgação científica	
	perfil (breve apresentação de uma personalidade)	
	classificado	
	horóscopo	
	obituário	
	boletim meteorológico	
comunicação comercial/disseminação	anúncios	
	cartazes	
	folhetos	
	ementa	

		rótulo	
		instruções/manuais de instruções	
		bula	
livro	não-ficção	(auto)biografia	
		crónica	
		ensaio	
		diário	
		prefácio/prólogo	
		dedicatória	
		autoajuda	
		relato de viagem	
		memória/testemunho	
		cartas	
		guia turístico	
	ficção	conto	
		fábula	
		épico	
		romance	
		lírico (poesia), letras de música	
		dramático	
	didático	manual**	** Excluir manuais de línguas.
		enciclopédia/atlas	
		livro de receitas	
		glossário	
académico		artigo	
		relatório	
		resumo/síntese	
		recensão crítica	
		proposta de projeto	
		dissertação/tese	
		ensaio	

político	discurso	compilar a partir das páginas dos partidos, de forma a garantir que os textos são direcionados para o público geral.
	moção/relatório	
	programa	
legal	lei	
	contrato	
	ofício/memorando/estatutos/edital	
religioso	oração	
	escrituras/salmos/epístolas	
	homilias	
	sermão/folheto/catequismo	

Anexo III – Distribuição total do número de textos e *tokens* por propósito, meio, tipo de documento e nível de complexidade

TIPOLOGIA	Muito Fácil	Fácil	Claro	(+)Complexo	TOTAL
comunicação pessoa					
diário	4	15	12	10	41
E-mail / carta pessoal	18	12	13	10	53
lista / agenda (de compras, de tarefas..)	29	14	14	10	67
bilhete	20	8	9	2	39
comunicação institucional/profissional					
carta / e-mail	16	17	13	8	54
relatório	1	2	16	11	30
comunicado de imprensa	2	10	10	10	32
comunicado/newsletter	5	9	8	10	32
página web (acerca de nós/missão...)	3	6	13	10	32
instruções	6	6	13	10	35
ata	0	4	21	6	31
comunicação social (jornais, revistas)					
editorial	6	17	12	10	45
notícia	27	18	10	10	65
reportagem	3	21	9	10	43
entrevista	44	8	13	11	76
artigo de opinião	10	10	13	11	44
artigo de divulgação científica	5	14	10	10	39
perfil	18	16	12	10	56
horóscopo	5	14	6	8	33
obituário	6	14	10	7	37
boletim meteorológico	5	10	10	10	35
comunicação comercial/disseminação					
anúncios	22	12	8	10	52
cartazes	13	9	6	10	38
folhetos	4	10	7	10	31
ementa	5	12	10	10	37
rótulo	4	14	11	10	39
instruções/manuais de instruções	4	11	9	10	34
bula	2	7	10	10	29
livro não-ficção					
(auto)biografia	10	13	11	15	49
crônica	6	8	6	10	30
ensaio	0	6	10	10	26
diário	12	15	9	11	47
prefácio/prólogo	5	13	10	10	38
dedicatória	9	8	11	3	31
autoajuda	6	17	11	10	44
relato de viagem	4	10	9	10	33
memória/testemunho	4	16	11	10	41
cartas	2	6	11	13	32
guia turístico	0	11	14	10	35
político					
discurso	0	3	13	14	30
moção/relatório	0	1	12	18	31
programa	0	4	7	19	30
legal					
lei	0	0	18	10	28
contrato	0	0	21	5	26

Anexo IV – Identificação de ULs e alinhamento de sentidos

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
adjetivo	DLP-antigo_1-f0dca	antigo		de outra época	10	relativo à Antiguidade, às épocas anteriores à queda do Império Romano	Pégaso, que era um cavalo da Grécia antiga , voava sobre os montes à beira do mar.	Grécia antiga		temporal
					1	que existiu em épocas muito anteriores à atual	O anão contava-lhe histórias do passado, histórias de mouros, guerreiros, navegadores, princesas e reis antigos .	Lisboa antiga		
				que existe há muito tempo	2	que tem muito tempo de existência	Mas também nos livros antigos de Ptolomeu e nos livros dos romanos está escrito que ninguém pode passar além do Bojador.			temporal
				que se deixou de usar	6	que já não se usa, que não é atual	Há também antigas linhas de comboio (onde o comboio já não passa) que agora são utilizadas para andar a pé ou para praticar outros desportos.	moeda antiga; forma antiga; caminho antigo		abstrato_ação
				que já deixou de trabalhar numa certa profissão	5	que deixou determinado estado, profissão ou cargo	Quem está a escrever-lhe esta carta é a sua antiga aluna do 1.º ciclo.			abstrato_ação

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
	DLP-branco_1-86241	branco		que tem a cor da neve, do leite	1	que tem a cor da neve, da cal, do leite	Sabes porque é que os pandas-gigantes são brancos e pretos?		rever 2. (ver a 4) + sentido para o que está pálido, lívido. rever 4. tendo em conta os exemplos seria um produto alimentar refinado. retirar 5? as combinações estão corretas? ou deveriam passar para o nível da entrada?	concreto
	DLP-leve_1-ca9bd	leve		que não é grave	6	que tem pouca importância, pouca gravidade	Acidente ainda causou dois feridos leves	ferido leve	sentido 9 não será o mesmo do 6? (cf. 8 de ligeiro)	abstrato_estad o
				que tem pouco peso	1	que tem pouco peso	Para andares com uma mochila mais leve prepara-a com o que realmente te faz falta para o dia seguinte.		rever 16 e 17 rever 18 e 19 rever 4 e 5	abstrato_estad o
				que é calmo, despreocupado, relaxado	13	que está aliviado, despreocupado	E tinha inveja das algas que baloiçavam ao sabor das correntes com um ar tão leve e feliz .	ar feliz		abstrato_estad o
				que é simples, pouco difícil	10	que é simples, despretenso, pouco carregado de ornamento	Nas férias, gosto de ler histórias leves e divertidas, nada de muito sério.	história leve		abstrato_estad o

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
					11	que trata de assunto banal, pouco profundo, superficial				
				que é suave ou pouco forte	8	que é pouco perceptível pelos sentidos; que é ténue, suave, ligeiro, delicado	Até ao dia em que ouviu um leve TOC, TOC!		rever 8 e 9	abstrato_estad o
				que é fino; que é pouco espesso	3	que é fino, pouco espesso	Utiliza agasalhos leves .	agasalho leve	rever exemplos e repensar definição.	abstrato_estad o
	DLP-seco_1-10262	seco		diz-se do alimento desidratado	5	que foi submetido a um processo de eliminação da humidade, garantindo desse modo a sua conservação	Mas ela sabia mais histórias, sabia a dos figos e contava todas as voltas que eles levavam até poderem ser chamados de figos secos .	figo seco alperce seco alho seco fruto seco		abstrato_estad o
				que não é fresco	3	que perdeu a humidade natural, a frescura	Caiu no escuro, por entre uma camada de fetos secos e musgos ainda húmidos.			abstrato_estad o
				diz-se da parte do corpo que não está molhada ou hidratada	7	diz-se de parte do corpo que não é suficientemente humedecida ou lubrificada pelas secreções	Senti a garganta seca e o coração a disparar.	garganta seca		abstrato_estad o
				que é caracterizado pela falta de água ou humidade	2	que é desprovido de humidade atmosférica	Afinal ainda estamos na estação seca e a chuva vai demorar algum tempo a chegar!	estação seca		abstrato_estad o
				que é brusco, áspero; que é pouco amável	16	que tem ou denota falta de sensibilidade, de ternura, de	- Cangurus, cucos e macacos - foi a resposta seca .	resposta seca		abstrato_estad o

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
						amabilidade ou de afabilidade				
advérbio	DLP- agora_1- 6142b	agora		usa-se para assinalar o presente	1	no tempo que está a decorrer; no momento presente ou nesta ocasião	Agora ouça o Carlos e verifique se ordenou as frases correctamente.			temporal
					3	deste momento em diante	Já deu aulas em muitas escolas diferentes, mas agora pensa que vai continuar na escola de Sintra.			
					4	depois disto; dadas estas circunstâncias ou em vista disto	Vamos agora relembrar o que é um adjetivo.			
					2	na época em que vivemos	Andávamos nós já na quarta classe - agora quarto ano - quando o Válter recebeu uma bicicleta como presente de anos.			
	DLP- mesmo_3- 31e9e	mesmo		usa-se para indicar intensidade, para reforçar algo	4	em grau elevado, de modo intenso, pleno	Amo-te mesmo, mesmo de verdade.			abstrato_estad o
				usa-se para indicar algo real, preciso	3	na realidade, em verdade, de modo efetivo	Voo mesmo a sério!			concreto
	DLP- sempre_1- 46cfa	sempre		usa-se para indicar todos os momentos ou todo o tempo de duração	1	em todos os momentos no tempo passado, presente ou futuro, indefinidamente; em todo o tempo	A minha sugestão é que façam uma lista de compras geral, para usarem todas as semanas e não andarem sempre a fazer uma nova.		ver 4; 5	temporal

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
					2	em todos os momentos e de modo constante, contínuo ou repetido, no passado, presente ou futuro; sem parar	Como mora perto da escola, anda sempre a pé.			abstrato_ ação
				usa-se para indicar algo que não se pode evitar	7	no momento considerado ainda e apesar de todas as circunstâncias contrárias	Eu era a prenda pelo teu décimo segundo aniversário e serias sempre tu a cuidar de mim.		rever a entrada toda (bom exemplo!) mas apenas porque está no condicional	abstrato_evento
	DLP-tarde_1-8f647	tarde		depois de um horário ideal ou oportuno	1	próximo do fim da noite	Mas o problema é que os filmes melhores normalmente passam tarde .			temporal
					2	próximo do fim da manhã				
					5	depois do momento considerado oportuno; além do tempo esperado ou conveniente; depois da ocasião própria			ver 1;2;4	
			mais tarde	algo que acontece depois; num momento posterior		em momento ulterior	Quando acabei o 12.º ano, fui para a universidade fazer um curso para mais tarde trabalhar na televisão.	mais tarde		temporal
nome	DLP-braco_1-795dd	braço		membro superior do corpo humano, do ombro à mão	1	cada um dos membros superiores do corpo humano, do ombro até à mão	Põem-nos no seu colo e protegem-nos com os braços .		ou ombro e cotovelo	humano concreto
					2	parte do membro superior				

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
						compreendida entre o ombro e o cotovelo				
	DLP- conta _1- ef0ad	conta		operação matemática	0	ato ou efeito de contar	És capaz de fazer as contas à quantidade de água gasta por toda a família?			abstrato_ ação
					1	operação aritmética para calcular ou determinar uma quantidade, um valor; resultado obtido nessa operação	- Não acertas nas contas , mas acertas nas montras.			
				resultado de uma compra	2	valor ou importância de uma despesa; total do que se deve	A demora no momento de pagar a conta .	pagar a conta		abstrato_evento abstrato_ ação
					3	fatura, documento onde está discriminado o que se deve				
				despesas de uma pessoa, normalmente mensais		-	Deixar as contas pagas.		sem equivalente	concreto humano
			perder a conta	deixar de saber a quantidade de algo		perder a noção da quantidade	Tocar nos dedos enquanto contas ajuda-te a não perder a conta .		uniformizar a definição com a que está na entrada perder	abstrato_estado
			fazer de conta	fingir ou imaginar algo		fingir	Uma das nossas brincadeiras era atar uma corda a dois paus e fazer de conta que eram duas vacas.			abstrato_ ação
			dar conta	reparar ou perceber algo		notar, reparar, perceber	O meu telemóvel pôs-se a fazer chamadas sem eu dar conta .			abstrato_ ação

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
			vezes sem conta	repetição		vezes sem fim	"Mais" era uma palavra que ela dizia vezes e vezes sem conta .			abstrato_estado abstrato_ação
			tomar conta	cuidar de algo ou alguém		encarregar-se; olhar por	Era ela que tomava conta dos onze filhos do moleiro.			abstrato_ação
	DLP- língua 1- e2f31	língua		forma de comunicação do ser humano	3	sistema de signos vocais, que podem ser transcritos graficamente, comum a um povo, a uma nação, a uma cultura e que constitui o seu instrumento de comunicação	O nosso trabalho é encontrar pessoas que falam línguas diferentes do português.			concreto humano abstrato_ação
				órgão móvel e comprido que está no interior da boca	1	órgão carnudo, musculoso, alongado e móvel, situado na cavidade bucal que serve para a degustação, para a deglutição e, no homem e em outros animais, para a articulação de sons	Pensam se devem ou não dar as mãos e deitar a língua de fora .	língua de fora		concreto humano animal
			língua oficial	aquela que é reconhecida por um Estado	-	a que é reconhecida estatutariamente por um Estado na qual são redigidos os textos oficiais por ele emanados	A língua portuguesa além de ser uma das línguas oficiais de Timor-Leste, é também falada em outros países do mundo.			humano concreto
			na ponta da língua	aquilo pronto a ser dito		sem hesitações nem dúvidas; muito bem	Tem sempre um facto científico na ponta da língua e adora partilhá-lo.			ação_estado humano

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
	DLP-pena_1-27e11	pena		sentimento de tristeza causado por algo ou alguém que se encontra numa situação negativa	1	sentimento suscitado por alguém ou alguma coisa que se encontra em estado de carência ou degradação	Martinho, que era bondoso e gostava de ajudar as pessoas mais pobres, ao ver o velho mendigo, ficou cheio de pena dele.			abstrato_estad o humano
					2	sentimento que se exprime por insatisfação e tristeza	Divertiam-se muito assim e, quando se separavam, sentiam sempre imensa pena .			
				aquilo que cobre, tapa o corpo das aves	1	órgão que cobre o corpo das aves, protegendo-o e permitindo a execução e orientação do voo, formado por uma haste flexível guarnecida de barbas	Um tinha penas , o outro, pelo.			concreto animal
			valer a pena	merecer o esforço ou o interesse		não merecer o esforço, a preocupação	É desagradável falar de dinheiro, não vale a pena .		valer de nível fácil. valer a pena – muito fácil	abstrato_event o
preposição	DLP-com_1-4d236	com		indica o meio, objeto ou instrumento com que se faz algo	9	meio, instrumento	Leve ao forno durante cerca de 30 minutos numa forma previamente untada com margarina e polvilhada com farinha.			concreto abstrato_ação
					14	matéria				
				indica o modo de fazer algo	6	modo	E nós sabemos que só a massa trabalhada com amor resulta.			concreto abstrato_ação
				indica posse	2	posse	Regressava à minha cidade, com uma mala e uma mochila.			concreto abstrato_estad o

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
				indica qualidade ou condição	3	qualidade ou estado, marcando a presença de uma característica inerente ou momentânea	Estava a ficar cansado e com fome.			concreto abstrato_estado
				indica relação entre algo ou alguém	1	companhia, presença física e relações entre pessoas	Está com problemas profissionais com os colegas e com os chefes.		ver parte II e III da entrada.	concreto
	DLP- desde _1- 9ba6a	desde		indica uma condição	1	desde que	Também são importantes os ralhetes - desde que justificados.			abstrato_estado
				indica o ponto no tempo ou espaço em que algo teve início	2	desde que	Esta semana assinalam-se dez anos desde que me mudei para Lisboa.			temporal locativo
				indica variedade entre elementos	II – 3	heterogeneidade dos elementos de um conjunto	As brincadeiras no banho, as musiquinhas que lhe cantares, aquilo que lhe ensinares - desde os livros aos carros - e até todas as vezes que a levares à bola: ela não se vai esquecer!			coletivo abstrato_estado
				indica o ponto no tempo ou espaço em que algo teve início	I – 1	ponto (no tempo ou no espaço) em que se inicia uma ação; processo ou estado de coisas	Uma abóbora que cá estava desde Outubro foi finalmente aberta.			temporal locativo
					I – 2	ponto de partida numa escala graduada	Os gatos não gostam de tomar banho e têm medo da água, pelo que devem ser habituados desde jovens.			

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
	DLP-durante 1-a0864-1	durante		indica a duração de um acontecimento ou evento	2	indica duração ou tempo contínuo em que decorre uma ação, um processo ou um estado	Lisboa tinha sido boa para mim durante quatro anos.			temporal abstrato_evento
				indica o intervalo de tempo de algum acontecimento ou evento	1	situa numa ação, um processo ou um estado num dado período ou espaço de tempo que lhe é simultâneo, sendo para parafraseável por «no tempo de», «no decorrer de»	Obrigada pelos votos de confiança e por terem aceitado o meu lema durante o confinamento.			temporal abstrato_ação
	DLP-entre 1-d38e8	entre		indica um intervalo entre duas ou mais opções	III – 2	quantidade intermédia no interior de dois limites indicados; valor aproximado	Um gato de pelo curto pode tomar entre 3 a 4 banhos por ano.			concreto
				indica relação entre duas ou mais pessoas ou coisas	I – 5	relação entre pessoas, traduzindo noções diversas como a reciprocidade, a união, a cooperação, a partilha	Só tem piada quando é para ser partilhada entre família e amigos.			
				indica alternativa ou escolha	I – 4	disjunção	Adoras programas entre amigos ou em família.			abstrato_ação
	DLP-cada 1-04a5d	cada		refere a parte individual de um todo	II	usa-se, por vezes, com valor de pronome: indica todo e qualquer elemento	É um desporto coletivo jogado entre duas equipas de 11 jogadores cada.			concreto

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
						de um grupo, considerado individualmente, sendo semanticamente equivalente a «cada um»				
	DLP- isso_1- 902d8	isso		refere algo que já foi dito ou mencionado	2	refere algo já mencionado ou a seguir explicitado	Eu sei que os <i>blogs</i> se querem leves e inspiradores mas não me identifico muito com isso .			abstrato_evento
			por isso	introduz uma conclusão		por esta razão	Já tenho, por isso , planos para entrar no mercado de trabalho, futuramente.			abstrato_ação
			para isso	introduz uma condição para conseguir algo		-	Para isso , temos de prestar atenção à forma como são produzidos os nossos alimentos.		sem equivalente	abstrato_ação
	DLP- quanto_1- 5c5ff	quanto		refere um número ou valor de objetos, pessoas ou ações	1	indica pessoas, coisas ou coisa com valor indeterminado de totalidade e apresenta função de sujeito ou de complemento, sendo parafraseável por «(todos) os que», «(todas) as que», «(tudo) o que» e podendo estar precedido ou não dos pronomes indefinidos todos, todas, tudo	Na sua vez, cada jogador lança o dado e desloca o seu peão colorido tantas casas quanto o dado indica. sempre seguindo a ordem crescente das casas.			concreto

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
	DLP-quant o_2-5c5ff				2	indica pessoas, coisas ou coisa com valor demonstrativo, apresentando maior ou menor proximidade em relação às pessoas do discurso e pode desempenhar função de sujeito ou de complemento, sendo parafraseável por «aquele(s) que», «aquela(s) que», «aquilo que» ou por «o(s) que», «a(s) que»	Um poeta inclina-se sobre os seus poemas: vinte foi quantos escreveu.			
					1	usa-se em frases interrogativas, diretas ou indiretas, em posição pré-nominal e com função de determinante e indica interrogação em relação ao número ou à quantidade de pessoas ou coisas, sendo parafraseável por «que número de», «que quantidade de»	Quantas vezes é que já saí de grandes superfícies de mãos a abanar?	quantas vezes		
			uns quantos	indica algumas pessoas ou coisas; uma quantidade indeterminada		algumas pessoas ou coisas	Fiz sopa para a semana e ainda uns quantos frascos de compota.			abstrato_estad o

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
	DLP- todo_1- ea5cb	todo		indica o total	1	qualquer elemento de entre um conjunto de semelhantes, com valor distributivo, e usa-se no singular	Fomos todos passar duas semanas ao Algarve.			coletivo abstrato_estado
					2	integridade de pessoa ou coisa, com valor adjetival, e usa-se no singular, omitindo-se o artigo definido, quando seguido de pronome pessoal	Naquele momento aquele pequeno ser é a nossa vida toda .			
			ao todo			por junto; no total	Ao todo estão cá cerca de 80 pessoas.		sem equivalente	
			de uma vez por todas	por última vez e de forma definitiva		-	Vem para que, de uma vez por todas , o que sinto possa fazer sentido.		é locução adverbial	concreto
verbo	DLP- bater_1- 3c5b2	bater		dar toques ou pancadas com alguma parte do corpo ou com um objeto	7	dar um ou mais pequenos toques ou pancadas em	Certo dia, o pai encontrou Dinis a bater numa parede com o nariz.			abstrato_ação
					8	dar pequenas pancadas com nós dos dedos ou com qualquer objeto numa porta, para pedir entrada	A tua mãe bateu à porta do quarto.	bater à porta	7,8,9 iguais?	
					9	dar várias pancadas com as mãos ou com um instrumento numa superfície, num objeto ou fazê-la embater	Outras gostam de fazer barulho na noite de passagem de ano, batendo com as tampas das panelas para afugentar o ano velho.		ver 3 e 4	

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
				chocar ou fazer chocar uma coisa em outra		contra algum obstáculo				
					5	fazer chocar ou chocar uma coisa com outra, com força	A chuva bate furiosamente na janela, o vento agita as cortinas.		o mesmo?	abstrato_ação
					6	ir, um automóvel, contra outro ou contra um obstáculo	Ele teve um acidente: bateu de frente.			
			bater (as) palmas	chocar as palmas da mãos uma contra a outra para demonstrar agrado, para aplaudir		dar com a palma de uma mão na outra repetidamente para exprimir aplauso ou para mostrar alegria	Ela sentia-se tão feliz que começou a bater palmas de alegria			abstrato_ação humano
	DLP-casar_1-adcfd	casar		unir ou unirem-se através do casamento	2	unir-se a alguém através do casamento	As três acabaram por casar com resineiros.		ver 1	humano abstrato_ação
					3	promover o casamento de alguém	Podes casar com a minha filha, podes ser rei quando eu morrer, podes ser tudo o que tu quiseres!		ver citação de 3	abstrato_evento
	DLP-sair_1-21552	sair		ir a algum lugar por diversão	4	deixar temporariamente o local onde se habita para tomar parte em atividades de caráter social ou lúdico	Excelente para sair à noite com os amigos.	sair à noite	o mesmo?	humano abstrato_ação abstrato_evento
					21	ir para fora de casa partilhando com alguém uma diversão ou outra atividade social				

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
				deixar um lugar	1	abandonar o local onde se está; passar do interior de um espaço fechado para o exterior	Depois de a minha mãe sair do quarto, aconteceu uma coisa estranha.			abstrato_ação locativo
					2	abandonar o local onde se está com o objetivo de ir fazer alguma coisa	Ao sair para a escola a mãe disse-lhe: -Volta depressa porque há muita roupa para lavar.		ver 14, 15 e 16 ver 23 e 24	
					3	deixar o local onde se trabalha ou estuda, concluída a tarefa diária	Tenho saído do escritório todos os dias pelas onze horas da noite.			
					5	abandonar, um meio de transporte, a estação de embarque, rumo ao seu destino	Se algum barco sai hoje ao mar, afunda-se!			
			sair do forno	passar de fora para dentro	1	passar do interior de um espaço fechado para o exterior	Já agora, todos os dias, temos pão quentinho a sair do forno .	sair do forno	É uma combinatória? >> coocorrente privilegiado com o sentido de algo fresco, recentemente feito, apresentado. Exemplos online: – Uma novidade da @planetatangeri na acabada de sair do forno!	abstrato_ação locativo

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
									– Celeste lança novidades acabadas de sair do forno. – Cinco novos livros de cozinha acabados de sair do forno.	
	DLP-vender_1-1a19b	vender		trabalhar como vendedor	3	exercer a profissão de vendedor por conta própria ou por conta de outrem; fazer comércio de determinada mercadoria	Faz bolinhos que a nossa avó vende no mercado, todas as manhãs.			humano abstrato_estado
				trocar algo por dinheiro	1	passar alguma coisa para a posse de alguém, em troca de determinada quantia em dinheiro; fazer a venda de	As castanhas assadas vendem-se nas ruas das cidades portuguesas		ver 8 e 9	humano abstrato_ação
					2	ter, um estabelecimento, uma firma comercial, determinada mercadoria disponível para compra; ter para venda	É uma pequena loja que vende jornais e revistas.		1 e 2 não pode ser só uma definição?	
	DLP-perder	perder		ignorar onde algo ou alguém se encontra	4	ser comprado com facilidade pelo público; ser vendável	É o terceiro bolo mais vendido nas pastelarias.			abstrato_estado

POS	ID	Lema	Sublema	Definição	N.	DLP alinhamento	Exemplo extraído do Sketch Engine	Coocorrente	Observações	Traço(s) Semântico(s)
	1-03f95				14	não saber onde se encontra, onde pôs, deixou ou colocou	Colocou a mala da senhora no chão e perdeu o dinheiro.		ver exemplos em 16	
				ficar desorientado ou distraído		-	Ficou perdido no meio das letras.		sem equivalente equivalente em perder-se, enquanto verbo pronominal	abstrato_estad o
				chegar atrasado	12	não apanhar um veículo, um meio de transporte, por ter chegado atrasado	la perder o comboio, de certeza!		alterar 12 para incluir 13? alterar definição de 13 para algo com o sentido de "não estar presente"?	humano abstrato_ação
				ficar sem algo que se tinha	5	deixar de ter, de possuir, uma característica, qualidade, defeito; ficar privado de; ficar sem	O início era bom, mas depois perde o interesse.		ver 10	abstrato_event o
			a não perder	algo que deve ser visto ou feito		-	Para mim, há dois sítios a não perder em Portugal.		sem equivalente	abstrato_ação
			perder a conta	deixar de saber a quantidade de algo		deixar de saber quantos são	Tocar nos dedos enquanto contas ajuda-te a não perder a conta.		uniformizar a definição com a que está na entrada conta	abstrato_estad o

Anexo V – Entrada “bater” antes da revisão



Dicionário da Língua Portuguesa
Academia das Ciências de Lisboa



(<https://www.acad-ciencias.pt/>)



Procure uma palavra



bater

Entrada editada

verbo

1. dar pancadas, com as mãos ou com um objeto, um instrumento, numa pessoa ou num animal, a fim de magoar, castigar

SINÓNIMOS agredir; dar pancada; sovar

«O marido era uma pessoa muito agressiva, [...] quis-lhe bater, etc, mas enfim a coisa lá serenou.» (PF, p. 151)

«Os irmãos [...] batiam nele como em cão malhadiço.» (Aquilino, *Terras do Demo*, p. 36)

«O pai bateu-te, mas o pai é teu amigo...» (Nemésio, *Mau Tempo*, p. 69)

2. INFORMAL obter a vitória, causando a derrota do inimigo ou do adversário

SINÓNIMOS vencer

ANTÓNIMOS perder

Os portugueses bateram os castelhanos em batalhas importantes.

3. INFORMAL ser melhor ou superior; ultrapassar em qualidade, perícia, capacidade

Este vinho bate o do ano passado em sabor e odor.

Em organização ninguém o bate.

4. ultrapassar

bater o recorde

Os números relativos ao crescimento económico bateram os níveis atingidos no ano passado.

5. fazer chocar ou chocar uma coisa com outra, com força

SINÓNIMOS embater

A chuva bate nos vidros.

«Elias Sobral bate com os pés no chão, a desentorpecer as pernas.» (J. L. Vieira, *Domingos Xavier*, p. 87)

«Leva as mãos ao chão, bate com o peito no joelho da outra perna.» (J. L. Vieira, *Domingos Xavier*, p. 170)

«penedias onde o mar batia e se espalhava branco e fino» (M. G. Azambuja, *Bárbara*, p. 282)

6. ir, um automóvel, contra outro ou contra um obstáculo

SINÓNIMOS chocar

«Tenho carta, já há seis aninhos [...]. Nunca bati, já me bateram duas vezes» (PF, p. 109)

7. dar um ou mais pequenos toques ou pancadas em

«Um espectador sentado na bancada de cima, [...] bateu-lhe levemente no ombro: — Pst... Oiça lá, cavalheiro!» (Nemésio, *Mau Tempo*, p. 445)

8. dar pequenas pancadas com nós dos dedos ou com qualquer objeto numa porta, para pedir entrada

bater à porta

Não deve entrar sem bater.

«Bateram à minha porta, / Fui abrir, não vi ninguém.» (M. Bandeira, *Obras Poéticas*, p. 364)

«Mioma foi bater à porta de Glorinhas. Bateu, chamou, ninguém lhe tornou resposta.» (Aquilino, *Terras do Demo*, p. 347)

9. dar várias pancadas com as mãos ou com um instrumento numa superfície, num objeto ou fazê-la embater contra algum obstáculo

bater os tapetes; bater as almofadas; bater a massa

10. misturar ou agitar vigorosamente com um instrumento adequado

bater ovos; bater as claras

«Depois, bate-se tudo, com açúcar, como uma gemada.» (J. L. Vieira, *Domingos Xavier*, p. 17)

11. agitar ou agitar-se com um ritmo regular, por vezes, com rapidez e energia

«um grande corvo amedrontado bateu as asas, elevou-se dum carvalho, grasnando» (Aquilino, *Terras do Demo*, p. 323)

«Demoro-me assim de pé, coração a bater desordenado» (M. T. Horta, *Ambas as Mãos*, p. 117)

12. assinalar ou ser assinalada uma hora, através do som das badaladas

«ainda o sino de santo Antão não bateu o meio-dia» (Aquilino, *Terras do Demo*, p. 251)

- «meia-noite a bater no relógio da Universidade, e eu, ligeiro como um gamo, a pegar na capa e a saltar para a rua!» (T. Coelho, *In Illo Tempore*, p. 241)
13. incidir fortemente sobre
- A luz desse candeeiro bate-me nos olhos.*
- «Um clarão de luz ardente caía sobre a cidade, vinha por detrás e batia-me nas costas» (H. Helder, *Passos em Volta*, p. 69)
- «o sol batia-lhe de chapa sobre o corpo» (M. T. Horta, *Ambas as Mãos*, p. 20)
- «a luz do sol bate-lhe na cara.» (J. L. Vieira, *Domingos Xavier*, p. 65)
14. atingir determinado ponto no espaço
- SINÓNIMOS tocar
- Daqui a pouco bates com a cabeça no teto.*
- A saia bate-lhe no joelho.*
15. atingir determinada situação, estado; chegar a
- «Ao nono toiro, a animação da praça batia o auge.» (Nemésio, *Mau Tempo*, p. 449)
16. INFORMAL percorrer um espaço, uma área, à procura de alguma coisa ou de alguém
- SINÓNIMOS explorar; inspecionar; percorrer
- Batemos a cidade toda em busca de um restaurante aberto.*
17. INFORMAL usar uma dada peça de vestuário com muita frequência, no dia a dia
- Bateu aquela saia quase durante um ano.*
- Comprei um fato melhor e outro para bater.*
18. GÍRIA jogar
- bater uma bisca; bater uma sueca*
19. dançar
- «Ao anoitecer [...] chamava-se o bombo, os ferrinhos, a gaita ou a viola. E iam bater a chula à porta de S. João até que dessem a voz de ceia.» (Aquilino, *Terras do Demo*, p. 284)
- «os chibantes armavam-se das castanholas [...] e batiam um fandango que até do mais fundinho do povo as raparigas vinham ver.» (Aquilino, *Terras do Demo*, p. 21)
20. GÍRIA fazer efeito, uma substância narcótica
21. BRASIL fazer-se sentir
- SINÓNIMOS chegar; manifestar-se
- De repente, bateu uma saudade enorme da minha terra.*
22. BRASIL, INFORMAL roubar
23. BRASIL chegar ou parar em determinado local, depois de uma deslocação longa
24. DESPORTO fazer uma batida
- SINÓNIMOS caçar

«Homem de seiscentos diabos aquele padre Francisco, tão escoteiro nas suas obrigações de pároco como, pelos montes, a bater as lebres!» (Aquilino, *Terras do Demo*, p. 211)

bater a asa

INFORMAL ir-se embora apressadamente

SINÓNIMOS fugir

bater a boa porta

encontrar a pessoa, o organismo, a instituição certa ou indicada

bater a bola baixo

ser humilde, submisso, não levantar problemas, objeções

bater a bota

GÍRIA morrer

«Puseram-lhe a extrema-unção e estive a bater a bota.» (Aquilino, *Terras do Demo*, p. 70)

bater a canastra

BRASIL, INFORMAL morrer

bater à máquina

dactilografar

bater ao/no coração de alguém

despertar determinado sentimento

bater aos pontos

ser muito melhor ou ganhar com grande distância de pontuação

bater a pala

fazer continência

bater (com) a porta

fechá-la com muita força, fazendo estrondo

«ouvía arredar cadeiras e alguém entrar para a casa de banho e bater a porta com estrondo.» (M. S. Lourenço, *Desequilibrista*, p. 109)

bater à porta certa

o mesmo que bater a boa porta

bater à porta de alguém

1. pedir ajuda
2. acontecer-lhe, suceder-lhe

«a fortuna não tardará a bater-te à porta...» (A. Redol, *Avieiros*, p. 300)

bater a sola

INFORMAL pôr-se em fuga

SINÓNIMOS fugir

bater a tamanquinha

dançar

«Quem a vira solteira, benzia-se. Na terra não nascera outra mais amiga de luxar e bater a tamanquinha em bailes e romarias.» (Aquilino, *Terras do Demo*, p. 15)

bater a todas as portas

recorrer a todas as ajudas possíveis

bater certo

ser exato

bater com a cabeça nas paredes

INFORMAL ficar desorientado, sem saber o que fazer

bater com a língua nos dentes

revelar um segredo

bater com a porta

ir-se embora, mostrando irritação

bater com o nariz na porta

encontrar fechada uma porta que se julgava estar ou vir a estar aberta

bater em retirada

fugir

bater mato/moita

dar pancadas na vegetação para obrigar os animais a mostrarem-se

bater na mesma tecla

repetir o que já se disse; voltar outra vez ao mesmo

bater no ceguinho

INFORMAL agredir ou acusar quem já não se pode defender

bater (com a mão) no peito

mostrar-se culpado e arrependido

bater o calcanhar

dançar

«Nas noites de sábado punha-se o trabalho e atafais para um canto, e batia-se o calcanhar.» (Aquilino, *Terras do Demo*, p. 62)

bater o compasso

marcar o ritmo com as mãos, os pés ou um instrumento

bater o dente

chocar os dentes de cima com os de baixo, por frio ou por medo

bater (a) o ferrolho

INFORMAL ir a casa de alguém

«a fulaninha já sabia que tal dia que tinha lá a malta a bater-lhe ao ferrolho.»
(PF, p. 1151)

bater o pé

INFORMAL insistir, com teimosia

SINÓNIMOS teimar

bater o ponto

ser o fulcro da questão

«Poderemos conseguir que ele convença o universo da sua fidalguia, sem se apossar dos vínculos de D. Mafalda? Aqui é que bate o ponto.» (Camilo, *Estrelas Propícias*, p. 137)

«aí é que bate o ponto!» (T. Coelho, *In Illo Tempore*, p. 279)

bater o(s) queixo(s)

o mesmo que bater o dente

bater palmas

dar com a palma de uma mão na outra repetidamente para exprimir aplauso ou para mostrar alegria

«as pessoas gostam e batem palmas...» (PF, p. 21)

bater uma foto/uma chapa

tirar uma fotografia

bater uma moeda

fabricá-la, imprimindo o cunho

SINÓNIMOS cunhar

bater uma soneca

INFORMAL dormir

bater um papo

BRASIL, INFORMAL conversar

ir bater (com os costados)

ir parar

não bater bem (da bola)

não ter juízo; ser desequilibrado

SINÓNIMOS não regular bem

bater bem

estar em seu perfeito juízo

ETIMOLOGIA Do latim *battuĕre*

Anexo VI – Entrada “bater” após revisão e edição

bater

verbo

1. dar pancadas ou golpes em **SINÓNIMOS** golpear
Ela bateu na mesa com força para chamar a atenção.
Os músicos bateram nos tambores com energia.
2. agredir fisicamente **SINÓNIMOS** espancar; sovar; surrar
«Os irmãos [...] batiam nele como em cão malhadiço» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 36)
3. dar um ou mais pequenos toques ou batidas em
«Um espectador sentado na bancada de cima, [...] bateu-lhe levemente no ombro: — Pst... Oiça lá, cavalheiro!» (Vitorino Nemésio, *Mau Tempo*, p. 445)
«Bateram à minha porta, / Fui abrir, não vi ninguém.» (Manuel Bandeira, *Obras Poéticas*, 1956, p. 364)
4. golpear repetidamente algo para remover sujidade, poeira ou auxiliar na limpeza
bater os tapetes; bater as almofadas
5. imprimir um ritmo regular, por vezes, com rapidez e energia
«um grande corvo amedrontado bateu as asas, elevou-se dum carvalhiço, grasnando» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 323)
«Demoro-me assim de pé, coração a bater desordenado.» (Maria Teresa Horta, *Ambas as Mãos*, 1970, p. 117)
6. misturar ou agitar vigorosamente com um instrumento adequado
bater ovos; bater as claras; bater a massa
«Depois, bate-se tudo, com açúcar, como uma gemada.» (José Luandino Vieira, *A Vida Verdadeira de Domingos Xavier*, 1974, p. 17)
7. fazer chocar ou chocar uma coisa com outra **SINÓNIMOS** embater
A chuva bate nos vidros.
«Elias Sobral bate com os pés no chão, a desentorpecer as pernas.» (José Luandino Vieira, *A Vida Verdadeira de Domingos Xavier*, 1974, p. 87)
«Leva as mãos ao chão, bate com o peito no joelho da outra perna.» (José Luandino Vieira, *A Vida Verdadeira de Domingos Xavier*, 1974, p. 170)
«penedias onde o mar batia e se espalhava branco e fino.» (Maria da Graça Azambuja, *Bárbara Casanova*, 1955, p. 282)
8. ir de encontro a, geralmente de forma violenta ou acidental **SINÓNIMOS** chocar; embater
As ondas batem no rochedo.
«Tenho carta, já há seis aninhos [...]. Nunca bati, já me bateram duas vezes» (PF, 109)
9. percorrer (espaço, zona) à procura de algo ou alguém **SINÓNIMOS** explorar; percorrer
Batemos a cidade toda em busca de um restaurante aberto.
10. vibrar sonoramente **SINÓNIMOS** dar; soar; tocar
«ainda o sino de santo Antão não bateu o meio-dia.» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 215)
«meia-noite a bater no relógio da Universidade, e eu, ligeiro como um gamo, a pegar na capa e a saltar para a rua!» (Henrique Trindade Coelho, *Estudantes, Lentes e Futricas*, 1943, p. 241)
11. incidir (raio luminoso, vento) fortemente sobre
A luz desse candeeiro bate-me nos olhos.
«Um clarão de luz ardente caía sobre a cidade, vinha por detrás e batia-me nas costas.» (Herberto Helder, *Os Passos em Volta*, p. 69)
«o sol batia-lhe de chapa sobre o corpo.» (Maria Teresa Horta, *Ambas as Mãos sobre o Corpo*, 1970, p. 20)

- «a luz do sol bate-lhe na cara.» (José Luandino Vieira, *A Vida Verdadeira de Domingos Xavier*, 1974, p. 65)
12. atingir determinado ponto no espaço **SINÓNIMOS** tocar
Daqui a pouco bates com a cabeça no teto.
A saia bate-lhe no joelho.
13. atingir determinado valor ou cifra **SINÓNIMOS** alcançar
 «Bolsa de Lisboa bate máximo de um ano.» (José Manuel Rocha, Público,)
 «Mas, nos grandes centros, a mexeriquece bate o auge.» (João Bigotte Chorão, *João de Araújo Correia – um clássico contemporâneo*, 1986, p. 89)
14. obter a vitória sobre (adversário, inimigo) **SINÓNIMOS** derrotar; vencer **ANTÓNIMOS** perder
Os portugueses bateram os castelhanos em batalhas importantes.
15. ser melhor ou superior **SINÓNIMOS** superar; ultrapassar
bater o recorde
Este vinho bate o do ano passado em sabor e odor.
Em organização ninguém o bate.
16. fabricar (moeda, medalha, etc.) imprimindo o cunho **SINÓNIMOS** cunhar
17. fechar (porta, tampa, etc.) com força
 «ouvira arredar cadeiras e alguém entrar para a casa de banho e bater a porta com estrondo.» (M. S. Lourenço, *O Desequilibrista*, 1960, p. 109)
18. despertar determinado sentimento **SINÓNIMOS** manifestar-se
bater ao/no coração de alguém
De repente, bateu uma saudade enorme da minha terra.
19. fazer com que a caça se levante ou corra; fazer uma batida **SINÓNIMOS** caçar
 «Homem de seiscentos diabos aquele padre Francisco, tão escoteiro nas suas obrigações de pároco como, pelos montes, a bater as lebres!» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 211)
20. **DESPORTO** chutar a bola num lance de bola parada
 «Ronaldo bateu forte um livre lateral, Guilherme deixou a bola escapar entre as pernas e Ramsey apareceu a confirmar o golo, ainda que tenha ficado a sensação de que a bola entraria.» (Mais Futebol, *Liga Campeões: Ramsey «tira» golo a Ronaldo, Douglas apura Juventus*, 06/11/2019)
21. **GÍRIA** sentir o efeito de (bebida alcoólica, substância narcótica)
22. **GÍRIA** jogar
bater uma bisca; bater uma sueca
23. **DESUSADO** escrever à máquina **SINÓNIMOS** datilografar
bater à máquina; bater uma notícia
24. **DESUSADO** usar (peça de vestuário) com muita frequência
Bateu aquela saia quase durante um ano.
Comprei um fato melhor e outro para bater.
26. **BRASIL, INFORMAL** roubar **SINÓNIMOS** surripiar
bater a carteira
27. **BRASIL** chegar ou parar em determinado local, depois de uma deslocação longa
28. **BRASIL** dirigir-se apressadamente

bater a asa

INFORMAL ir-se embora apressadamente **SINÓNIMOS** fugir

bater a boa porta

encontrar a pessoa, o organismo, a instituição certa ou indicada **SINÓNIMOS** bater à porta certa

bater a bola baixo

ser humilde, submisso, não levantar problemas, objeções

bater a bota

GÍRIA morrer **SINÓNIMOS** bater a canastra

«Puseram-lhe a extrema-unção e esteve a bater a bota.» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 70)

bater a canastra

BRASIL, INFORMAL morrer **SINÓNIMOS** bater a bota

bater aos pontos

ser muito melhor ou ganhar com grande distância de pontuação

bater a pala

fazer continência

bater à porta certa

encontrar a pessoa, o organismo, a instituição certa ou indicada **SINÓNIMOS** bater a boa porta

bater à porta de alguém

1. pedir ajuda

2. acontecer-lhe, suceder-lhe

«a fortuna não tardará a bater-te à porta...» (Alves Redol, *Avieiros*, 1942, p. 300)

bater a sola

INFORMAL pôr-se em fuga **SINÓNIMOS** fugir

bater a tamanquinha

dançar

«Quem a vira solteira, benzia-se. Na terra não nascera outra mais amiga de luxar e bater a tamanquinha em bailes e romarias.» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 15)

bater a todas as portas

recorrer a todas as ajudas possíveis

bater certo

ser exato

bater com a cabeça na(s) parede(s)

INFORMAL ficar desorientado, sem saber o que fazer

bater com a língua nos dentes

revelar um segredo

bater com a porta

ir-se embora, mostrando irritação

bater com o nariz na porta

encontrar fechada uma porta que se julgava estar ou vir a estar aberta

bater em retirada

fugir

bater mato/moita

dar pancadas na vegetação para obrigar os animais a mostrarem-se

bater na mesma tecla

repetir o que já se disse; voltar outra vez ao mesmo

bater no ceguinho

INFORMAL agredir ou acusar quem já não se pode defender

bater (com a mão) no peito

mostrar-se culpado e arrependido

bater o calcanhar

dançar **SINÓNIMOS** bater a tamanquinha

«Nas noites de sábado punha-se o trabalho e atafais para um canto, e batia-se o calcanhar» (Aquilino Ribeiro, *Terras do Demo*, 1946, p. 62)

bater o compasso

marcar o ritmo com as mãos, os pés ou um instrumento

bater o dente

chocar os dentes de cima com os de baixo, por frio ou por medo

bater (a) o ferrolho

INFORMAL ir a casa de alguém

«a fulaninha já sabia que tal dia que tinha lá a malta a bater-lhe ao ferrolho.» (PF, 1151)

bater o pé

INFORMAL insistir, com teimosia **SINÓNIMOS** teimar

bater o ponto

ser o fulcro da questão

«Poderemos conseguir que ele convença o universo da sua fidalguia, sem se apossar dos vínculos de D. Mafalda? Aqui é que bate o ponto.» (Camilo Castelo Branco, *Estrelas Propícias*, 1965, p. 137)

«aí é que bate o ponto!» (Henrique Trindade Coelho, *In Illo Tempore*, 1943, p. 279)

bater o(s) queixo(s)

chocar os dentes de cima com os de baixo, por frio ou por medo **SINÓNIMOS** bater o dente

bater (as) palmas

dar com a palma de uma mão na outra repetidamente para exprimir aplauso ou para mostrar alegria

bater uma foto/uma chapa

INFORMAL tirar uma fotografia

bater uma soneca

INFORMAL dormir

ir bater (com os costados)

ir parar a

não bater bem (da bola)

não ter juízo; ser desequilibrado **SINÓNIMOS** não regular bem

bater um papo

BRASIL, INFORMAL conversar

bater bem

estar em seu perfeito juízo

ETIMOLOGIA Do latim clássico *battuĕre* ‘bater, brigar’, pelo latim vulgar *battĕre* ‘dar pancadas em’