

Extração e descrição do vocabulário fundamental para o ensino do Português Europeu como língua de acolhimento

Cláudia Rafaela Bação Castro

Dissertação

de Mestrado em Ciências da Linguagem: Terminologia e Gestão de Informação de Especialidade

Versão corrigida e melhorada após defesa pública

Janeiro, 2024

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Ciências da Linguagem: Terminologia e Gestão de Informação de Especialidade

Realizada sob a orientação científica da
Professora Doutora Raquel Amaro

Macte animo sic itur ad Astra

AGRADECIMENTOS

Em primeiro lugar, gostava de agradecer à minha família pelo seu apoio e compreensão durante esta etapa da minha vida.

Em segundo lugar, queria agradecer aos meus amigos, e colegas de mestrado, Alexandre, Alice e Joana, pela disponibilidade, pela ajuda e pelos seus incentivos durante o Mestrado e durante a redação desta dissertação. Um obrigada também a todos os meus amigos que me acompanharam durante este percurso.

Gostava de agradecer à minha orientadora, a Professora Raquel Amaro, por toda a ajuda que me deu, tanto na orientação, mas também na revisão do meu trabalho científico e na minha escrita.

Por fim, gostava de agradecer ao projeto EXPRESSI pela bolsa de investigação (licenciado), do Centro de Linguística da Universidade NOVA de Lisboa (CLUNL), UIDB/03213/2020, que me providenciou uma oportunidade de trabalhar numa instituição como o CLUNL.

Resumo

O tema do trabalho desenvolvido nesta dissertação é o vocabulário fundamental do ensino do português como língua de acolhimento. O português como língua de acolhimento é uma iniciativa de auxílio ao acolhimento e integração de pessoas migrantes em situações de acolhimento. Contudo, existem indícios que apontam para o subdesenvolvimento dos materiais de ensino do português como língua de acolhimento. O trabalho executado no âmbito do projeto EXPRIMI consiste na apreciação do estado dos materiais de ensino do português como língua estrangeira e na obtenção do vocabulário fundamental do ramo do acolhimento e da integração.

A apreciação dos materiais de ensino do português como língua estrangeira faz-se através da comparação de *corpora*, e a definição do vocabulário fundamental advém da extração de informação linguística vinda do *corpus* de foco.

Os resultados das etapas descritas trazem a oportunidade de mostrar o uso do material recolhido através da proposta de um recurso lexicográfico.

PALAVRAS-CHAVE: Vocabulário fundamental; Português língua de acolhimento; Linguística de *corpus*; Extração de informação linguística; Lexicografia;

Abstract

The theme of the work developed in this dissertation is the fundamental vocabulary of Portuguese teaching as a host language. Portuguese as a host language is an initiative to assist in the reception and integration of migrant people in host situations. However, there is evidence that teaching materials for Portuguese as a host language are underdeveloped.

The work carried out within the project EXPRIMI consists of assessing the state of the teaching materials of Portuguese as a foreign language and obtaining the fundamental vocabulary on the field of hosting and integration.

The evaluation of the teaching materials of Portuguese as a foreign language is done through the comparison of *corpora*, and the definition of the fundamental vocabulary comes from the extraction of linguistic information from the focus *corpus*.

The results of the steps described provide the opportunity to show the use of the collected material through the proposal of a lexicographic resource.

KEYWORDS: Fundamental vocabulary; Portuguese host language; *Corpus* linguistics; Linguistic information extraction; Lexicography;

LISTA DE SIGLAS E ABREVIATURAS

QECR – Quadro Europeu Comum de Referência para as Línguas

TBLT – *Task Based Language Teaching*

PLE - Português como Língua Estrangeira

ACM – Alto Comissariado de Migrações

IEFP - Instituto de Emprego e Formação Profissional

PLA – Português Língua de Acolhimento

CNQ - Catálogo Nacional de Qualificações

CQL – *Corpus Query Language*

OCR – *Optical Character Recognition*

SEF - Serviço de Estrangeiros e Fronteiras

CPR - Conselho Português para os Refugiados

Conteúdo

1.Introdução	1
1.1. Integração no projeto EXPRESSI	4
1.2. Questão de investigação: contexto.....	5
1.2.1 Léxico Fundamental	5
1.2.2. O português como língua de acolhimento	7
2. Enquadramento teórico	10
2.1. Linguística de corpus.....	10
2.1.1 Ferramentas de trabalho lexicográfico: Sketch Engine	12
2.2. Lexicologia e lexicografia	13
2.2.1. Elementos de um recurso lexical	14
2.2.2. Ferramentas de trabalho lexicográfico: Lexonomy	15
3. Metodologia de trabalho	16
3.1. Trabalho de corpora	17
3.1.1. Corpus MIGRANTE.PT	17
3.1.2. Corpus MATERIAIS_PLE_2023.....	18
4. Extração e descrição de vocabulário.....	22
4.1. Extração de vocabulário.....	22
4.2. Análise dos resultados automáticos	23
4.3. Organização em campos semânticos	30
4.3.1. Definição dos campos relevantes.....	31
4.3.2. Os campos semânticos	33
4.4. Proposta de glossário.....	35
4.4.1. Público-alvo	36
4.4.2. Micro e macroestrutura no Lexonomy.....	36
4.4.3. Questões relevantes para a compilação	40
4.4.4. Exemplos de entradas	41
5. Considerações finais	43
6. Referências.....	46
Lista de Figuras.....	48

Anexo 1: Listas de unidades por campo semântico	49
Anexo 2: Descrição do glossário (Lexonomy)	53
Anexo 3: Entradas do glossário Português Língua de Acolhimento	54

1.Introdução

O conhecimento de uma segunda língua é uma mais-valia no mundo atual.

O ensino do português como língua estrangeira pode abrir portas para pessoas migrantes encontrarem novas oportunidades em Portugal. A aprendizagem de uma língua pode acontecer tanto para enriquecimento pessoal, como também pode ser uma ferramenta essencial à integração de pessoas migrantes, para encontrarem melhores circunstâncias em Portugal. O uso eficaz do português é fundamental para a comunicação de uma pessoa migrante para assuntos essenciais do seu quotidiano, bem como no caminho para a cidadania no país de acolhimento.

O ensino de língua estrangeira tem várias vertentes, sendo o ensino de uma língua como língua de acolhimento uma das situações intrinsecamente ligadas com este ensino, e neste caso o ensino português como língua de acolhimento dá-se no contexto do ensino a pessoas migrantes que procuram refúgio em Portugal.

O português na função de língua de acolhimento é um assunto moderno com desenvolvimento recente no seu ensino e na construção de materiais de ensino. Todas as pessoas migrantes têm direito à língua. Para exercerem os seus direitos existem cursos de português como língua de acolhimento. As aulas de língua estrangeira têm como objetivo dar conhecimento linguístico explícito, para promover o uso da língua em questão ao aluno. O aluno, por sua vez, usaria este conhecimento na integração da sociedade estrangeira. Em Portugal, as entidades disponíveis para o ensino de português para estrangeiros são muitas vezes as universidades ao oferecerem cursos de português como língua estrangeira, cursos estruturados para o ensino de língua estrangeira em ambiente educativo.

Tal como em qualquer ensino de língua, os estudantes do português como língua de acolhimento necessitam de material para adquirir conhecimento linguístico, útil para o seu quotidiano.

Dentro da área do ensino de língua estrangeira, o contexto da lexicologia remete para a identificação e organização de vocabulário para o ensino. No contexto de ensino de português como língua estrangeira, o ensino de vocabulário tem normalmente como base o chamado vocabulário fundamental (Bacelar do Nascimento *te al.* 1984; Quadro

européu comum de referência para as línguas (QECR), 2001; Referencial Camões PLE, 2017), i.e., um conjunto que abrange as unidades lexicais mais comuns no português europeu.

O material de ensino do português precisa do vocabulário certo. Este material deverá ser adequado ao nível de proficiência linguística e adequado à comunidade linguística alvo. Devido à situação extraordinária da comunidade linguística das pessoas migrantes em situação de acolhimento, existe a questão de o material de ensino ser adequado às suas circunstâncias. O propósito do ensino de língua de acolhimento é permitir a independência de mediadores a pessoas migrantes através do domínio da língua. O vocabulário fundamental, embora um recurso elementar ao ensino do léxico geral, pode não cobrir as exigências linguísticas específicas de uma situação como o acolhimento de pessoas migrantes. Isto pode causar falhas na utilização do português como língua de acolhimento nas circunstâncias descritas, devido ao carácter geral do vocabulário ensinado.

As situações de acolhimento, em particular, envolvem uma junção de áreas à volta da recepção de pessoas em circunstâncias vulneráveis, estas áreas vão para além do cumprimento dos níveis de proficiência ou do ensino do vocabulário dito fundamental ou nuclear (Grosso, 2010).

É pertinente que o vocabulário ensinado a pessoas migrantes em situação de acolhimento acompanhe as necessidades dos aprendentes, beneficiando estes de conteúdo que dão apoio à sua situação. Existe então, a questão de o material ser adequado às circunstâncias, ou seja, permitir o domínio da língua e a independência a pessoas migrantes.

Um estudo recente (Amaro *et al.*, 2022) mostra que os materiais usados para o ensino do português em situações de acolhimento podem não ser adequados a esta perspetiva. De acordo com o artigo, os materiais didáticos em questão podem não conseguir acompanhar a situação de uma pessoa migrante em situação de acolhimento, nem ser úteis para a sua realidade. No estudo referido, foi atestado que os aprendentes de português como língua de acolhimento podem ter dificuldade em preencher formulários, compreender documentos ou executar tarefas básicas essenciais para a sua estadia e integração devido a uma educação pobre em vocabulário dirigido às suas necessidades específicas. Isto demonstra que existem razões para afirmar que pessoas

migrantes em situação de acolhimento não estão a receber uma educação de português como língua estrangeira correspondente às suas necessidades. O conteúdo das aulas de português como língua estrangeira pode não estar a preparar os alunos em situação de acolhimento na exposição da terminologia importante que pode ser encontrada na sua realidade, levando a assumir um ensino pouco eficaz na integração especial para as pessoas migrantes em acolhimento. Outro ponto em Amaro *et al.* (2022) é a possibilidade de existir discrepâncias entre o vocabulário fundamental no ensino do português e o léxico utilizado em documentos destinados a situações de acolhimento e integração de migrantes.

O ensino de uma língua vocacionado para o acolhimento pode despoletar mais interatividade entre a comunidade linguística do país de acolhimento e a comunidade de migrantes em acolhimento, bem como ajudar o aprendente a obter melhor o conhecimento linguístico.

A obtenção do vocabulário fiel à experiência do acolhimento no ensino de língua seria uma mais-valia para pessoas migrantes em situação de acolhimento, na medida em que permitiria contribuir para a sua integração e autonomia. Esta ideia também tem o potencial de expandir o vocabulário fundamental das aulas de português como língua de acolhimento dando mais ferramentas comunicacionais aos alunos a frequentarem as mesmas.

Neste contexto, o trabalho desenvolvido aspira realizar os seguintes objetivos:

- Determinar, com base em dados reais, qual o vocabulário fundamental para o processo de acolhimento;
- Determinar qual o vocabulário específico à situação de acolhimento que está ausente em materiais de português língua estrangeira;
- Organizar o vocabulário em campos semânticos.
- Apresentar o vocabulário de acolhimento num recurso lexicográfico.

1.1. Integração no projeto EXPRIMI

O trabalho proposto foi feito no âmbito do projeto EXPRIMI – *Exploração preliminar de corpus para ensino inclusivo da língua de acolhimento baseado em tarefas*¹, ao abrigo de uma bolsa de investigação (licenciado), do Centro de Linguística da Universidade NOVA de Lisboa, UIDB/03213/2020.

O projeto EXPRIMI, em colaboração com o Conselho Português para os Refugiados, tem como foco “a investigação preliminar sobre o ensino-aprendizagem do português como língua de acolhimento”, com três metas a alcançar. Estas são²:

“i) métodos de identificação, extração e descrição do vocabulário fundamental para a aprendizagem de língua de acolhimento, utilizando ferramentas e procedimentos computacionais;”

“ii) análise semiautomática de textos e seus respetivos contextos com vista à identificação e extração de traços e critérios para monitorização de tendências/preconceitos nos textos e no uso de língua;”

“iii) métodos de conceção e desenvolvimento de materiais de ensino de língua baseado em tarefas para o português, beneficiando o contexto específico de aprendizagem e as necessidades comunicacionais de integração.”

Considerando os três pontos principais, o projeto delineou as suas vertentes de investigação: recolher o vocabulário fundamental para o ensino do português como língua acolhimento com ferramentas computacionais e aplicar este vocabulário fundamental ao ensino do português; estudar o uso enviesado de língua através da análise de textos; explorar métodos para desenvolver materiais de ensino TBLT (*task based language teaching*) em português para situações de aprendizagem de língua de acolhimento.

¹ <https://clunl.fcsh.unl.pt/projetos/projetos-curso/exprimi/>

² <https://clunl.fcsh.unl.pt/projetos/projetos-curso/exprimi/>

O trabalho desenvolvido contribui para o primeiro objetivo do projeto EXPRESSI. A extração e seleção do vocabulário fundamental para o ensino-aprendizagem do português como língua de acolhimento, obtido através da linguística de corpus.

A proposta para a utilização do vocabulário faz uso da lexicografia na forma de um recurso lexicográfico. O estudo aproveita as disciplinas da linguística de *corpus*, para a fase do tratamento de dados e a identificação e extração do léxico fundamental, e das disciplinas da lexicologia e lexicografia para o tratamento lexicográfico.

Através da linguística de *corpus*, o objetivo é recolher o léxico a partir de textos de situações de comunicação reais com vista ao acolhimento de pessoas migrantes, e em materiais de ensino do Português como Língua Estrangeira (PLE), agregados em *corpora* específicos, que incluem diversos géneros e tipos de textos. A partir destes dados, foi extraído o léxico aqui tratado com o intuito de perceber quais as necessidades em termos de vocabulário de pessoas migrantes, bem como a identificação do vocabulário em falta nos materiais didáticos do PLE. Para além dos resultados em termos de listas de vocabulário, a linguística de *corpus* permite ainda a extração e observação de informação linguística a partir de *corpora* de grandes dimensões.

O objetivo seguinte é propor o tratamento lexicográfico do léxico relevante bem como a demonstração da utilidade do recurso proposto, um glossário, por meio de entradas lexicográficas.

1.2. Questão de investigação: contexto

1.2.1 Léxico Fundamental

O conceito de *léxico fundamental* resulta de longa investigação existente sobre métodos teóricos, estatísticos e de linguística de *corpus* para compilar listas de palavras que acomodam vocabulário frequente – recolhidas usando as frequências de ocorrências e informação acerca do género de textos de *corpora* monitores e acerca do *vocabulário disponível*, isto é, vocabulário temático com menor probabilidade de ocorrência nos *corpora* monitores, mas reconhecido como essencial à comunicação (Quadro europeu comum de referência para as línguas (QECR), 2001; Referencial Camões PLE, 2017).

Existe um léxico fundamental já elaborado para o português, o *Português Fundamental* (Bacelar do Nascimento *et al.* 1984), com o objetivo de fornecer aos professores de

língua conjuntos de palavras que sejam úteis ou adequados para o ensino/aprendizagem de língua segunda/estrangeira. Este projeto foi inspirado em outras criações de léxico fundamental, particularmente as obras do Francês Fundamental e o Espanhol Fundamental. Consistiu na reunião de uma amostra variada da população portuguesa (igualdade no número dos participantes em termos de género, geografia, situação económica) para poder observar a comunicação entre falantes nativos sobre os temas conversacionais e léxico envolvido no português fundamental. Foi relatado pelos autores (cf. p.25) os constrangimentos que o regime político da altura impôs para a recolha de léxico para o Português Fundamental.

No entanto, para além de poder estar já desatualizado, é necessário perceber se este léxico cobre as necessidades de comunicação de pessoas migrantes em contexto de acolhimento, com vista à sua integração e autonomia.

Além disso, estudos recentes indicam que ainda há questões específicas a serem abordadas ao usar esses léxicos para identificação de palavras complexas – arcaísmos, empréstimos, expressões multipalavras (Tomé Cortejo, 2015).

A questão da aprendizagem do português como língua de acolhimento com vista à questão da integração das pessoas migrantes na sociedade é abordada durante o decorrer deste trabalho. A aprendizagem de uma língua em situação de acolhimento ou mudança permanente pode requerer, por exemplo, a instrução da língua de acordo com a linguagem dos textos oficiais ou incluir os termos legais dentro do léxico lecionado, sendo isto mais vantajoso na medida em que permite adquirir conhecimento linguístico correspondente à realidade e necessidades comunicacionais da pessoa migrante. Este tipo de trabalho já existente sobre material de ensino de português como língua estrangeira para efeitos de acolhimento averigua a utilidade dos conteúdos pedagógicos destinados às comunidades no ensino do português.

As pessoas migrantes em situação de acolhimento podem ser consideradas um grupo de estudantes misto, podendo haver necessidades e níveis de proficiência diferentes entre estes. O vocabulário institucional relacionado com a situação de acolhimento tem um registo formal, devido à sua origem legislativa.

A dificuldade do vocabulário fundamental numa situação de acolhimento seria a introdução e adaptação de léxico formal e complexo para a compreensão dos alunos de português como língua de acolhimento.

O estudo do vocabulário fundamental é executado pela pesquisa de léxico, tomando variáveis estatísticas em consideração, como a aquisição das unidades lexicais mais frequentes. Através da observação do material dedicado ao ensino de língua atual, podemos determinar o estado dos materiais de aprendizagem do português relativamente a este público de aprendentes. Esta investigação pode ser feita pela seleção e análise de documentos oficiais direcionados a migrantes e respetivo levantamento de unidades lexicais. Estas unidades serão consideradas pela sua importância contextual.

Posteriormente, é possível fazer uma comparação com os materiais dirigidos para ensino-aprendizagem do português como língua de acolhimento e verificar a presença das unidades consideradas pertinentes na fase de análise anterior.

Ao identificar o estado atual dos materiais destinados ao ensino do português como língua segunda, particularmente em situações de acolhimento e integração (Celso & Marques, 2019), será possível melhorar os recursos pedagógicos em questão.

1.2.2. O português como língua de acolhimento

O objetivo do ensino do português como língua de acolhimento consiste em assistir na integração de pessoas migrantes através da formação no português.

Neste trabalho, faz-se a distinção entre acolhimento e integração: Acolhimento diz respeito à chegada das pessoas migrantes e às suas necessidades linguísticas imediatas; e integração diz respeito às necessidades linguísticas vindas da estadia no país de chegada. As necessidades linguísticas podem dizer respeito a situações médicas (como ler bulas de medicamentos e receitas médicas), burocráticas ou jurídicas (preenchimento de formulários e documentação com dados pessoais), sociais e culturais (compreender as infraestruturas dos transportes e comunicação informal), familiares, etc. Pelo lado jurídico do português como língua de acolhimento, a competência linguística também é considerada um requisito para a aquisição de estatutos de residência, como o de longa duração e residência permanente ou para o processo de obtenção de nacionalidade na República Portuguesa. Cada estatuto referido pede, no mínimo, um nível de proficiência A2 para satisfazer os requisitos do governo português. Estes requisitos pedem que o requerente comprove a sua proficiência no português

Europeu, facto que prova a importância da proficiência do português para todas as pessoas migrantes com intenções de obterem cidadania, para além das necessidades sociais.

“Conhecer a língua do país de acolhimento não é apenas uma condição necessária e indispensável para se ser autónomo, é também, e sobretudo, condição de desenvolvimento pessoal, familiar, cultural e profissional. O seu desconhecimento constitui uma desigualdade que fragiliza as pessoas, tornando-as dependentes e, por consequência, mais vulneráveis. Poder aprender a língua do país de acolhimento é poder adquirir os meios de comunicar, interagir, compreender, defender-se, confrontar-se com uma outra cultura e outros códigos, é poder escolher abrir-se aos outros.” - (Portaria n.º 1262/2009 de 15 de Outubro, Diário da república, 1ª série - N.º 200 – 15 de Outubro de 2009)

A partir da portaria nº183/2020, há a criação dos cursos de “Português Língua de Acolhimento” específicos à situação de acolhimento de migrantes.

O português como língua de acolhimento, especificamente, é promovido por entidades como o Alto Comissariado para as Migrações (ACM) e o Instituto de Emprego e Formação Profissional (IEFP), que dispõem de programas para o ensino de português como língua não materna e de acolhimento, tal como o “Programa PPT (Português Para Todos)” do IEPF e os cursos de Português Língua de Acolhimento (PLA) do ACM. Os cursos estão disponíveis em agrupamentos escolares em todo o país, espaços escolhidos para acolher o projeto das instituições IEPF e ACM. Para auxiliar os cursos de PLA, o IEPF fornece materiais específicos. Estes materiais são: o “caderno de formação”; o “guia para o ensino do português enquanto língua de acolhimento no contexto da educação não formal (ENF)” e duas plataformas *online* “Plataforma de Português Online” e “Plataforma de Recursos Pedagógicos”.

De modo geral, pode dizer-se que há interesse na integração de pessoas migrantes em situação de acolhimento e que esse interesse também se demonstra no desenvolvimento de cursos de ensino do português específicos à sua situação. Os

programas acima foram descritos como tendo preocupação pelas necessidades do aluno em situação de acolhimento, refletindo-se esta preocupação nas tarefas direcionadas aos aprendentes do português.

O ensino do PLA, como todas as vertentes de ensino de língua estrangeira, precisa de material relativo a vocabulário específico, que permite à pessoa migrante o contacto com o ambiente social do país de acolhimento, o ambiente cultural, educação e formação profissional. Os materiais existentes, reflexo do esforço do ACM e das outras entidades, estão detalhadas na página da organização³.

O vocabulário já existente nos livros de ensino das entidades acima pode ser extraído e compilado ao lado do vocabulário mais comum no processo de acolhimento, permitindo a criação de um glossário com o vocabulário fundamental do português como língua de acolhimento.

O percurso de formação de PLA do ACM e do IEFP faz ligação com o Quadro Europeu Comum de Referência para as Línguas (QECR)⁴, bem como com o Catálogo Nacional de Qualificações (CNQ) para formar os conteúdos de vocabulário para estes cursos. A organização do vocabulário dos cursos em questão consegue observar-se por tópicos - família, escola, lazer, anatomia, cores, alimentação, cultura, etc. - distribuídos por quatro níveis de proficiência de língua. (QUADRO EUROPEU COMUM DE REFERÊNCIA PARA AS LÍNGUAS – Aprendizagem, ensino, avaliação, 2001: 83-84).

Para além do QECR e do QNC, os materiais “O Português para Falantes de Outras Línguas – O Utilizador Elementar no País de Acolhimento”, “Português para Falantes de Outras Línguas – O Utilizador Elementar no País de Acolhimento - Sugestões de Atividades e Exercícios” e “O Português para Falantes de Outras Línguas – O Utilizador Independente no País de Acolhimento” serviram como base para a elaboração do caderno de formação e também das plataformas do IEFP. Estes materiais servem assim como base para o currículo dos níveis de proficiência A1, A2, B1 e B2.

³ <https://www.acm.gov.pt/pt/-/como-posso-frequentar-um-curso-de-lingua-portuguesa-para-estrangeiros->

⁴ Conselho da Europa (2001). QUADRO EUROPEU COMUM DE REFERÊNCIA PARA AS LÍNGUAS – Aprendizagem, ensino, avaliação. *Colecção: Perspectivas actuais/Educação*. Edições ASA. https://www.dge.mec.pt/sites/default/files/Basico/Documentos/quadro_europeu_comum_referencia.pdf

Os recursos disponibilizados pelo ACM foram alvo preferencial de observação por serem dirigidos diretamente a pessoas migrantes. Por exemplo, uma análise às atividades exemplo sugeridas revela o objetivo dos materiais em questão em ensinar português virado para o uso quotidiano da pessoa migrante.

A experiência de acolhimento em específico está rodeada de burocracia, termos técnicos, legislativos e institucionais possivelmente não referidos no QECR e pouco vistos no currículo de educação dos cursos PLA. A ausência deste vocabulário nos cursos de PLE com vista ao acolhimento e integração tem o risco de não ajudar os formandos em situação de acolhimento no caminho da integração. A falta de meios para comunicar pode levar a dificuldades em estabelecer uma vida em Portugal às pessoas migrantes nesta situação vulnerável. Grosso (2010) argumenta que a língua de acolhimento engloba necessidades comunicativas profissionais, de direitos e de cidadania. Estes domínios ainda não foram confirmados nos materiais.

2. Enquadramento teórico

2.1. Linguística de *corpus*

A linguística de *corpus* é usada neste trabalho como método de recolha de informação linguística. A linguística de *corpus* serviu para ter uma amostra da língua usada e atual dentro do contexto do projeto. Através desta amostra foi possível recolher vocabulário relativo a situações de acolhimento. A área da linguística de *corpus* é responsável pelo uso de dados autênticos, o *corpus*, para análise da língua. Um *corpus* é uma coleção de textos autênticos representativos de uma língua. Um *corpus* contém uma amostra de léxico usado em contextos reais (Sinclair, 1991; Tognini-Bonelli, 2001; Hunston, 2002). A metodologia de análise de *corpus* está relacionada com cálculos de frequência e estatística de tokens e lemas⁵, um dos processos da análise de informação linguística conseguida pela utilização de *corpora*.

Recorre-se ao uso de *corpora* como fonte de informação para criar recursos relacionados com a língua. Obras como dicionários, glossários e vocabulários de

⁵ De acordo com o Dicionário de Termos Linguísticos, um lema é “Forma gráfica escolhida convencionalmente como vedeta de um dicionário ou de um léxico. MULLER (1974)”
<http://www.portaldalinguaportuguesa.org/?action=terminology&act=view&id=3168>

especialidade são criados a partir da compilação de *corpora* para obter as unidades mais frequentes na área, para serem incluídos como nomenclatura.

O uso de *corpora* também é notado na extração de vocabulário para o ensino de língua. De acordo com Cotos (2017), a linguística de *corpus* obtém dados linguísticos autênticos a serem aproveitados para objetivos pedagógicos. Estes, por sua vez, podem ser utilizados na construção de vocabulário específico, contribuindo para um desenvolvimento linguístico eficaz e para o desenvolvimento de competências práticas e adequadas a cada situação.

A linguística de *corpus* teve uma grande atualização no último século. A era digital tornou mais fácil usar a linguística de *corpus* como metodologia de investigação linguística por várias razões:

- Pela evolução do *world wide web* e da sua capacidade de guardar e curar informação, tornando o acesso a dados reais muito mais fácil. Pela disponibilidade cada vez maior de informação *on line*. Atualmente é uma opção muito acessível recolher textos para investigação e observar o léxico em tempo real.
- Pela evolução de ferramentas digitais para criação de *corpora* e das ferramentas de processamento de língua natural.

Com os mecanismos tecnológicos disponíveis atualmente, como a criação de *corpora* ou a extração automática de informação linguística (por exemplo, em ferramentas como o Sketch Engine), a criação de recursos lexicográficos torna-se uma tarefa mais facilitada. O projeto EXPRIMI utiliza a linguística de *corpus* na forma de exploração de *corpora* já existente para obter vocabulário fundamental para o português como língua de acolhimento e usar o vocabulário obtido para criar um recurso lexicográfico. Dentro do recurso lexicográfico, faz-se o uso de *corpora* mais uma vez para retirar um exemplo de um lema ou expressão em contexto real. A linguística de *corpus* não vai servir só para extração de informação linguística, mas também como base de recolha de exemplos genuínos de utilização do vocabulário específico.

2.1.1 Ferramentas de trabalho lexicográfico: Sketch Engine

Uma das ferramentas digitais disponíveis para criar e explorar *corpora* é o Sketch Engine (<http://www.sketchengine.eu>).

O Sketch Engine permite a compilação e a exploração de *corpora*. Como ferramenta, o Sketch Engine prepara os textos computacionalmente, processa o material carregado com a tokenização automática de sequências; com a etiquetagem da categoria gramatical dos *tokens*⁶ automaticamente e da sua lematização. A ferramenta permite também fazer comparação entre dois *corpora* para motivos de análise linguística. O processamento das sequências de texto em *tokens* etiquetados e lematizados permite o uso da *Corpus Query Language* (CQL) para pesquisar instâncias linguísticas. Isto permite mais precisão na pesquisa de concordâncias de um lema.

A ferramenta Sketch Engine suporta várias unidades. Já foram referidos lemas e *tokens*, mas o Sketch Engine tem um lema que inclui classe de palavras: o *lempos* = *lemma* + *Part-of-Speech*. Dentro da funcionalidade de *lempos*, o Sketch Engine recolhe o lema acompanhado de categoria gramatical e desta maneira levanta-se dois tipos de dados em apenas uma pesquisa. A atribuição categoria gramatical é dada automaticamente pela ferramenta.

As funcionalidades da ferramenta Sketch Engine mais relevantes para o trabalho desenvolvido foram o Word Sketch, as *Keywords*, as concordâncias e os Good Dictionary Examples.

O Word Sketch consiste na representação da distribuição de um dado lema ou forma, mostrando as unidades mais frequentes à volta deste, o que ajuda a definir o contexto em que o lema surge (Kilgarrieff, 2014). A funcionalidade da extração de *keywords*, que providência listas com as unidades mais relevantes de um dado *corpus*; A concordância é uma funcionalidade que procura todas as formas de um dado *corpus*, mostrando a palavra pesquisada ao centro e uma janela de 5 ou mais elementos de cada lado. Em termos de extração de linhas de concordância, a funcionalidade *Good Dictionary Examples* também foi utilizada, consistindo na pesquisa de contextos em frases simples, sem pronomes ou anáforas, para o propósito de as usar como exemplos em recursos do tipo dicionário (ver Kilgarrieff, 2014).

⁶ Em Natural Language Processing (NLP), um *token* é a unidade mais pequena de significado.

O trabalho desenvolvido tirou partido do Word Sketch pela sua capacidade de análise distribucional; das *keywords* para extração de vocabulário fundamental; da concordância para consultar o contexto dos resultados e classificá-los de acordo com os campos semânticos definidos e dos *good dictionary examples* para a seleção de exemplos das entradas do recurso lexicográfico proposto.

2.2. Lexicologia e lexicografia

O léxico é, em primeiro lugar, imaterial. O léxico, enquanto parte da capacidade humana da linguagem, tem uma base cognitiva (Klein, 2001:8674). No entanto, este objeto imaterial, para servir o seu objetivo comunicacional, tem de ser partilhado por uma comunidade linguística.

O léxico é composto por várias formas com significado semântico e gramatical, chamadas neste trabalho como unidades lexicais. A lexicologia, uma vertente da linguística, estuda as formas com significado e a representação desta vertente da língua. Considera-se um conjunto de unidades lexicais um vocabulário. O Dicionário de Termos Linguísticos descreve vocabulário como⁷:

Lista exhaustiva das palavras de um corpus. O vocabulário distingue-se do léxico que é entendido como o inventário de todas as lexias de um dado estado de língua. Denomina-se ainda vocabulário um dicionário constituído pelos vocábulos mais frequentes da língua corrente, definidos sucintamente, seguidos por vezes do seu equivalente numa língua antiga ou estrangeira. Guilbert (1971)

Há várias maneiras de classificar um vocabulário dentro da Lexicologia. Existe vocabulário comum, vocabulário técnico, vocabulário fundamental, vocabulário de base, vocabulário de especialidade, entre outros.

A seguir existe a Lexicografia, cujo objeto de são os recursos lexicográficos, como dicionários e glossários. Um dicionário é um recurso lexicográfico e área de interesse da Lexicografia, construído pelo um aglomerado de unidades lexicais definidas morfossemanticamente. Um glossário já é um tipo de recurso que engloba unidades

⁷ <http://www.portaldalinguaportuguesa.org/?action=terminology&act=view&id=3275>

lexicais específicas, com definições mais simples. A definição do Dicionário de Termos Linguísticos consta: “Denomina-se glossário um dicionário que contém sob forma de simples definições (ou traduções) as significações das palavras raras ou pouco conhecidas.” Dubois *et al.* (1973)⁸

Estes recursos têm em comum uma macroestrutura e uma microestrutura. Uma microestrutura encarrega-se de definir a unidade lexical, definindo-a com informação morfológica, semântica e outros tipos. A microestrutura funciona a nível da entrada lexicográfica, e a macroestrutura funciona a nível de tudo o resto de um recurso. A macroestrutura engloba o resto do recurso, como capa, introdução, instruções de uso e referências.

Como na linguística de *corpus*, os recursos lexicográficos também mudaram com a era digital, o que permitiu a transformação dos recursos lexicográficos de cópias físicas para um novo formato: cópias eletrónicas (L’Homme, 2014; Lino, 2018). Ferramentas como o Lexonomy, uma plataforma para criação e edição de dicionários é apenas um exemplo da digitalização da lexicografia.

O recurso lexicográfico elaborado dentro deste trabalho é uma proposta de glossário digital para o português como língua de acolhimento, dirigido a pessoas migrantes em situação de acolhimento.

2.2.1. Elementos de um recurso lexical

Microestrutura

A entrada lexicográfica abrange a informação lexical que precisa de ser transmitida num espaço curto. A representação de uma unidade lexical depende dos objetivos do recurso e, tipicamente, requer uma microestrutura capaz de acolher informação de várias áreas da linguística (Hanks, 2003:78): ortografia, classe de palavras, morfologia, fonologia e semântica lexical. Logo, uma entrada lexicográfica contém, num espaço reduzido, um vasto leque de elementos linguísticos.

Uma entrada lexicográfica tipicamente contém: o lema; uma ou mais definições do ou dos significados associados o lema em questão; informação que põe a forma em contexto, mais conhecida como exemplos de uso, a categoria sintática, a etimologia e a

⁸ <http://www.portaldalinguaportuguesa.org/?action=terminology&act=view&id=3147>

fonética. Dependendo do tipo de recurso, também existem módulos para relações com outras formas (sinonímia, antonímia, etc.), ou em certas ocasiões, até pode conter imagens ou diagramas.

A era digital também afetou a lexicografia, causando um impacto na microestrutura, sendo possível agora incluir efeitos sonoros e visuais numa entrada lexicográfica. A origem de outro formato para recursos lexicográficos abriu novas maneiras de incluir conhecimento linguístico. O dicionário eletrónico infopédia, por exemplo, já acompanha os verbetes com faixas de áudio.

A proposta de glossário vai conter o formato de microestrutura descrito acima, com a unidade lexical, categoria gramatical, duas aceções em português e em inglês e exemplo de uso. Vai haver a inclusão do campo semântico a que a vedeta pertence e a inclusão de palavras sinónimas e palavras relacionadas.

Macroestrutura

A macroestrutura de um recurso lexicográfico diz respeito ao recurso no seu todo. É a macroestrutura que organiza a nomenclatura e as referências do recurso, introduz o recurso lexicográfico e também enquadra a utilização do mesmo. A macroestrutura também serve para ordenar as entradas que constituem o corpo do recurso por ordem que pode ser hierárquica, alfabética, por tema ou qualquer outro tipo de organização de vocabulário.

Em termos de macroestrutura da proposta de glossário, o recurso é ordenado por campos semânticos. A intenção é agrupar o vocabulário por áreas importantes à experiência do migrante em situação de acolhimento e incorporar essa organização do vocabulário na macroestrutura da proposta.

2.2.2. Ferramentas de trabalho lexicográfico: Lexonomy

Para a realização da proposta de glossário foi escolhida a ferramenta digital Lexonomy. O Lexonomy (Měchura, M. 2017) é uma ferramenta eletrónica onde é possível produzir conteúdo lexicográfico. O Lexonomy está concebido para a criação e a edição de dicionários, permitindo a elaboração de uma nomenclatura de acordo com as exigências da investigação.

O Lexonomy disponibiliza dois modelos padrão de dicionário (dicionário monolíngue e bilingue), onde são acrescentadas entradas lexicográficas. O Lexonomy permite também configurar a macroestrutura, como selecionar a língua do recurso, tornar o recurso público ou privado e dar acesso a outros utilizadores ao recurso lexicográfico.

O Lexonomy também permite a definição da microestrutura. As entradas lexicográficas também têm modelos padrão como opção, mas a ferramenta também dá a oportunidade de personalizar uma entrada lexicográfica de acordo com as necessidades do utilizador.

Isto significa que o Lexonomy tem a flexibilidade para adicionar ou eliminar elementos, acrescentar mais que um elemento ou criar etiquetas, bem como outro tipo de informação necessária, dependendo da vontade do utilizador.

Uma vantagem de um recurso digital é a opção de conectar partes da microestrutura ou das entradas. As entradas lexicográficas construídas no Lexonomy podem conter ligações que remetem para outras entradas do mesmo recurso, através de uma hiperligação. A opção de estruturar ligações entre entradas para obter uma conexão de palavras para representar relações semânticas entre as unidades lexicais é uma mais-valia para o projeto, bem como consome menos tempo na busca da informação lexical complementar. Tudo está disponível na microestrutura da própria entrada lexicográfica. Para aceder ao Lexonomy existe um domínio digital gratuito (www.lexonomy.eu).

3. Metodologia de trabalho

Os passos delineados para atingir os resultados pretendidos foram os seguintes:

- Atualização do *corpus* “MATERIAIS_PLE_2023” como *corpus* de referência, um *corpus* de materiais de ensino do PLE, de modo a garantir um *corpus* com alguma dimensão e representativo dos materiais existentes.
- Comparação dos *corpora* na ferramenta digital Sketch Engine para extração de léxico específico. Este processo foi completado automaticamente pela ferramenta, através da funcionalidade 'Keyword extraction', que consiste na observação do *corpus* de foco, MIGRANTE, em comparação com o *corpus* de referência, MATERIAIS_PLE_2023.
- Recolha de listas de *lempos* únicos ao *corpus* de foco MIGRANTE.PT. A finalidade desta fase é adquirir conhecimento de unidades lexicais utilizadas em situações de

acolhimento para averiguar como adaptar o português para situações de vulnerabilidade. Ao conhecer o léxico utilizado nos processos burocráticos que lidam com situações de acolhimento, este conhecimento trará a oportunidade de compilar candidatos a elementos do vocabulário relevante para providenciar uma melhor compreensão das necessidades linguísticas de migrantes, que pode ajudar o português como língua estrangeira e torná-lo mais específico para migrantes em situações de acolhimento.

- Observação dos *lempos* únicos recolhidos para a elaboração de grupos semânticos com os mesmos.

- Agrupamento das unidades lexicais em grupos semânticos relacionados com o processo de acolhimento de acordo com vários recursos relacionados com a imigração, acolhimento e refúgio.

- Utilização das unidades lexicais e dos grupos semânticos para elaborar uma estrutura para um recurso lexicográfico e adaptação das unidades lexicais para entradas lexicográficas.

3.1. Trabalho de *corpora*

3.1.1. *Corpus* MIGRANTE.PT

Em 2021, o projeto EXPRIMI já havia compilado um *corpus* centrado no tema de migração e processo de acolhimento de migrantes. O *corpus* em questão, resultou da recolha de documentos legislativos, notícias, relatórios e outros textos relevantes ou expositivos do processo de migração e acolhimento de requerentes de asilo.

Com esta informação, foi montado o *corpus* MIGRANTE.PT⁹. O MIGRANTE.PT contém aproximadamente 1,5 milhões de *tokens*.

O *corpus* referido é o *corpus* de foco para a extração de informação linguística. O *corpus* de foco na linguística de *corpus* consiste no alvo de exploração linguística para adquirir elementos linguísticos como: expressões compostas candidatas a elementos do

⁹ Amaro, Raquel, Susana Correia & Matilde Gonçalves (2021). *MIGRANTE.PT. 2021*. ISLRN: 872-735-563-701-2. <https://clunl.fcsh.unl.pt/recursos-em-linha/corpora/migrante-pt/>; DOI: <https://doi.org/10.34619/ctho-dqqk>

vocabulário relevante; lemas com altos níveis de ocorrência únicas ao *corpus* de foco ou em grande maioria no *corpus* de foco; concordâncias e exemplos de uso em contexto.

3.1.2. *Corpus* MATERIAIS_PLE_2023

A possibilidade de comparar *corpora* é o primeiro passo na investigação para a averiguação do vocabulário existente. É um passo muito importante no estudo das necessidades linguísticas de pessoas migrantes em situação vulnerável.

Para ser possível a extração de informação linguística num *corpus* de foco é necessário um *corpus* de referência para servir de comparação. Para servir o contexto do projeto foi criado um segundo *corpus* para fins específicos: o *corpus* MATERIAIS_PLE_2023.

O MATERIAIS_PLE_2023 foi compilado através da junção de manuais escolares e materiais de ensino do português, com o objetivo de variar em níveis de proficiência de A1 a C2. Este *corpus* fornece um ângulo importante à investigação, pois demonstra o input linguístico das aulas de português como língua estrangeira. É seguro supor que o input presente nos materiais das aulas de língua estrangeira seja uma percentagem do conhecimento explícito dos alunos de línguas. Este *corpus* tem como objetivo mostrar o vocabulário de ensino no contexto de sala de aula.

Contudo, o *corpus* MATERIAIS_PLE_2023 era insuficiente em tamanho, necessitando de uma atualização para servir adequadamente como *corpus* de referência. O *corpus* em questão não tinha a dimensão adequada para fazer a comparação com o *corpus* de foco. O MIGRANTE.PT contém aproximadamente 1 milhão e meio de *tokens*, algo não observado no MATERIAIS_PLE_2023 pré-atualização (com cerca de 435 0000 *tokens*).

O cruzamento de dados fornecidos pelos *corpora* conduz a investigação sobre o vocabulário necessário no acolhimento a pessoas migrantes. Este passo oferece a informação necessária para chegar ao vocabulário específico numa situação de acolhimento.

A primeira etapa foi recolher material para povoar o *corpus*, sendo referido anteriormente a sua insuficiência em termos de tamanho.

O material escolhido para a manutenção do *corpus* consistiu em manuais escolares, cadernos de exercícios, gramáticas do português para estrangeiros (para ensino ou para estudo independente) e faixas áudio de exercícios de pronúncia e compreensão,

complementares ao ensino de língua. Em termos das faixas áudio recolhidas, estas correspondiam maioritariamente em exercícios de diálogo e monólogos, com o ocasional exercício de repetição de som. As faixas de áudio recolhidas eram um auxílio para os manuais obtidos para o ensino da fonética do português. A fonética é um aspeto muito importante no ensino de língua e a sua inclusão tornou-se pertinente na recolha de materiais para investigação.

A próxima etapa foi a conversão dos materiais para texto simples.

A conversão para texto simples tem um aspeto diferente para cada tipo de material existente. Para as faixas de áudio recolhidas para o *corpus*, foi feita uma transcrição de áudio bruta. Esta transcrição foi automática, onde a funcionalidade “ditar” da ferramenta Microsoft Word transcreveu as faixas, seguida pela revisão da transcrição. A ferramenta em questão transcrevia o áudio, mas falhava na criação de frases porque falhava a colocar pontuação e, em certos casos, falhava também na ortografia. Considerando o funcionamento da ferramenta “ditar” a tarefa de transcrição envolveu adicionar pontuação mínima na criação de frases simples. Arranjos complexos, como por exemplo diálogo, foram processados com pontuação simples e construções simples. Ao todo foram incorporados sete documentos com transcrições de áudio provenientes dos CD dos manuais de ensino.

Para os materiais escritos em formato PDF (manuais, gramáticas e cadernos de exercícios) houve uma conversão para texto simples, dependendo do estado do ficheiro. Os manuais, cadernos de exercícios e gramáticas em formato PDF de boa qualidade e com leitura ótica de caracteres foram convertidos para texto simples num processo simples. No entanto, nos casos dos materiais em formato PDF onde não existia leitura ótica, foi necessária uma preparação dos mesmos para possibilitar a sua conversão. Para este efeito, foi feito uso de um sistema OCR (*Optical Character Recognition*), consistindo na transformação de documentos em vários formatos para reconhecimento de caracteres e leitura computacional. Este processo de OCR foi realizado sobre os materiais em formato de imagem, para poder ser possível a sua conversão para texto simples e a incorporação dos textos no *corpus*.

Para além da assistência de um programa de leitura ótica, também houve uma preparação dos materiais para a sua utilização: elementos como capa, prefácios, fichas técnicas e algumas introduções foram removidas para uma melhor interpretação dos

dados ao prevenir o mínimo de “ruído” possível. A grande parte do conteúdo a povoar o *corpus* é o núcleo do manual, como os índices, a gramática da língua, os textos, o vocabulário e os exercícios apresentados aos alunos.

Os passos descritos foram necessários como pré-processamento do material recolhido. A conversão para texto simples é imprescindível para o processo de compilação de *corpora*, pois é a partir do texto simples que ferramentas como o Sketch Engine são capazes de processar os textos.

Atualmente, as ferramentas de construção de *corpora* consomem menos tempo e recursos. A computadorização dos recursos linguísticos consegue facultar os resultados quantitativos básicos para investigação linguística. O Sketch Engine foi a ferramenta de análise de texto utilizada no trabalho de atualização de *corpora* para a investigação presente. O programa escolhido tem a capacidade de automatizar o processo de lematização e de tokenização das unidades lexicais, bem como consegue apresentar estatísticas sobre o *corpus* construído. As estatísticas apresentadas pelo Sketch Engine são uma boa fonte de informação para exploração, pois mostram resultados interessantes sobre o conteúdo que constitui o *corpus*.

Por exemplo, a frequência de ocorrência dos *tokens* consegue verificar quais unidades lexicais são mais comuns no *corpus* de foco e menos representadas no *corpus* de referência. Esta informação será útil na criação de entradas lexicográficas, pois apontará para as expressões não representadas nas aulas de português como língua de acolhimento e poderá criar uma oportunidade para adaptar os materiais com as unidades lexicais recolhidas.

A atualização do *corpus* MATERIAIS_PLE_2023 originou um *corpus* com 1 105 917 *tokens*. O *corpus* final aumentou significativamente de volume e, após a intervenção, conseguiu servir como *corpus* de referência para a investigação linguística que decorreu.

Anterior à sua atualização, o *corpus* MATERIAIS_PLE_2023 continha cerca de 435 000 *tokens*. Na fase de atualização do *corpus* foram adicionados mais 28 recursos de ensino de português como língua estrangeira. Sete destes recursos foram transcrições de áudio; cinco foram gramáticas, treze eram manuais de ensino e cinco foram cadernos de exercícios.

O *corpus* atualizado tem uma distribuição que cobre, ainda que de forma não equilibrada, todos os níveis de proficiência. A diversidade dos manuais didáticos em termos de proficiência vai desde o nível A1 ao nível C2. Todos os níveis de proficiência estão representados, existindo um volume maior de materiais de nível A1 e B1, com o nível C2 o nível de proficiência menos representada.

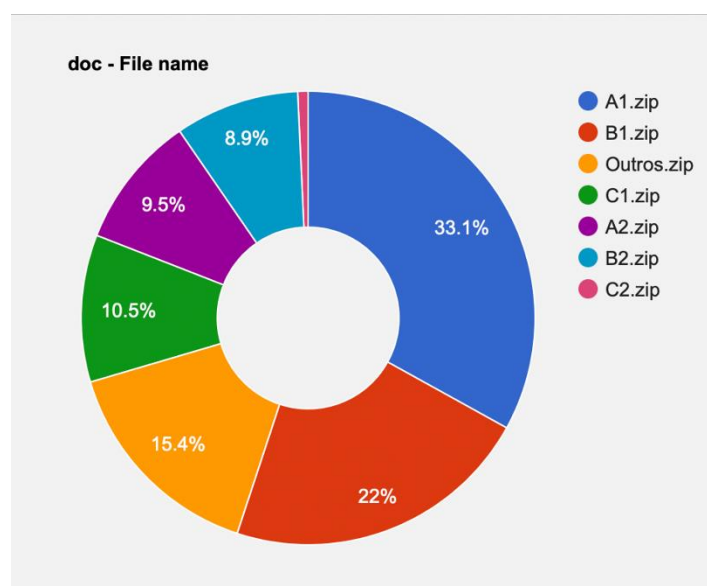


Figura 1: Composição do corpus MATERIAIS_PLE_2023 por nível de proficiência.

Attribute value	Token coverage ?
1 A1.zip	367,283 ...
2 B1.zip	244,257 ...
3 Outros.zip	171,518 ...
4 C1.zip	116,102 ...
5 A2.zip	105,476 ...
6 B2.zip	98,135 ...
7 C2.zip	8,177 ...

Figura 2: Número de tokens em cada nível de proficiência no corpus MATERIAIS_PLE_2023.

4. Extração e descrição de vocabulário

4.1. Extração de vocabulário

A funcionalidade da extração de *keywords* é capaz comparar os *tokens* dos dois *corpora* (foco e referência) ao usar o corpo de cada *corpus* na análise de ocorrências de frequência. A funcionalidade junta os *tokens* com a maior frequência presentes no *corpus* de foco, mas não representadas *corpus* de referência.

O objetivo desta etapa da metodologia é: extrair as unidades com mais ocorrência do *corpus* MIGRANTE.PT que se encontram em falta no *corpus* MATERIAIS_PLE_2023. Os *tokens* mais comuns no *corpus* MIGRANTE.PT são comparados com o número de ocorrências repetidas no *corpus* MATERIAIS_PLE_2023. Foi escolhido um levantamento de *lempos* para ter a categoria gramatical das unidades. A utilização de *lempos* na extração de *keywords* permite a recolha de lemas com a categoria gramatical que classifica as unidades.

Após o tratamento do material para atualização do *corpus* de referência e a sua incorporação na ferramenta Sketch Engine, vem a extração de informação linguística.

As unidades candidatas à nomenclatura do glossário foram obtidas através de listas produzidas pela extração de *Keywords*. Foram extraídas duas listas: uma de unidades simples e uma de unidades (ou expressões) multipalavra. Nestas listas, as expressões multipalavra e unidades simples mais comuns dentro do *corpus* MIGRANTE.PT são ordenadas pela frequência que ocorrem dentro do *corpus* de foco. As frequências de ocorrência destas *keywords* são equiparadas com o *corpus* MATERIAIS_PLE_2023, que por sua vez mostra o número de ocorrências no *corpus* de referência após a sua atualização.

A partir desta etapa, a extração de unidades do *corpus* MIGRANTE.PT, compôs-se o vocabulário de situações de acolhimento a ser utilizado na proposta de glossário do português como língua de acolhimento.

Com a comparação de *corpora* entre o MIGRANTE.PT e o MATERIAIS_PLE_2023 foi possível extrair uma lista de *keywords* com 1000 *lempos* simples; e uma lista de *keywords* com 1000 expressões multipalavra. Ao todo foram recolhidas 2000 unidades para a análise linguística.

4.2. Análise dos resultados automáticos

A etapa da análise dos dados extraídos inicia-se na recolha de listas de *keywords*: unidades simples e expressões multipalavra.

As listas são definidas pelos objetivos do projeto. O projeto quer obter todas as expressões únicas ou em grande maioria no *corpus* de foco. Para conseguir o pretendido, os parâmetros de pesquisa são definidos de acordo com o objetivo.

Os parâmetros de pesquisa das listagens foram construídos com o atributo de *lempos*, uma funcionalidade da ferramenta que permite obter lemas e categoria gramatical na mesma pesquisa. As listas foram delimitadas a 1000 *lempos* de unidades simples e 1000 expressões multipalavra por 1000 ser o resultado por defeito da funcionalidade de extração de *keywords*. No entanto, foi averiguado nas primeiras listas extraídas certas unidades lexicais, oriundas dos relatórios incorporados no *corpus* MIGRANTE.PT, causavam interferência com a exploração de dados linguísticos. Os relatórios não são documentos direcionados a pessoas migrantes nem servem a situação de acolhimento. Devido a esta interferência nas listas de *keywords*, foi criado um *sub-corpus* intitulado “MIG_REPORT_FREE”. A criação de um *subcorpus* excluiu os relatórios, o que ajudou a focar os resultados da pesquisa dentro do que era preciso recolher. O MATERIAIS_PLE_2023 foi utilizado na pesquisa como *corpus* de referência e o MIG_REPORT_FREE foi utilizado como *corpus* de foco. A partir destes parâmetros, a amostra recolhida do *corpus* MIGRANTE.PT foi criada automaticamente pelo Sketch Engine.

Na fase da análise dos *lempos* extraídos foram observados os seguintes fenómenos:

- Foi apreciado o facto de a ferramenta selecionar como unidades simples o que na realidade são expressões multipalavra. Este fenómeno acontece mais nas classes gramaticais de adjetivos e nomes.

Um exemplo deste fenómeno é o *lempos* simples “predial-adj”. A ferramenta processou a unidade isolada quando é mais útil se for usada numa expressão multipalavra. De acordo com as 17 ocorrências, 8 das ocorrências são de “caderneta predial”, 4 das ocorrências são de “matriz predial” e nas restantes ocorrências o adjetivo está sempre antecipado por “registo”.

Um exemplo real para um nome seria o *lempos* “encarregado-n”. Ao consultar as concordâncias do Sketch Engine do *lempos* em questão foi visto que “encarregado-n” ocorria sempre acompanhado por “de educação”. Ou seja, dentro do MIGRANTE.PT, esta unidade simples faz parte de uma expressão multipalavra. O contexto do *lempos* em questão é “encarregado de educação”. Tanto o *lempos* simples “encarregados-n” como a expressão multipalavra “encarregado de educação” estão presentes nas listas recolhidas e a expressão multipalavra foi selecionada para a o campo semântico “Educação”.

Outro exemplo do mesmo fenómeno, mas com um resultado ligeiramente diferente, é o *lempos* simples “alergénicos-n” que, na realidade, fazia parte da expressão “anti-alergénicos”. Ao olhar para as concordâncias foi possível percebê-lo, mas a expressão multipalavra “anti-alergénicos” não consta na lista de *keywords* multipalavra extraída. O que levou à adaptação da unidade lexical no campo semântico “Saúde”.

O mesmo aconteceu para o *lempos* “prescrição-n”, com 9 exemplos de ocorrência, onde aparece 4 vezes como “prescrição médica”, 2 vezes como “prescrição medicamentosa” e 3 vezes solto, mas sempre dentro do contexto de prescrição de medicamentos. Nenhuma destas expressões está presente na lista de expressões multipalavra. Este *lempos* foi adotado como “prescrição médica”, de acordo com a maioria dos exemplos de concordância (o *lempos* adotado está na lista do campo semântico “Saúde” no anexo 1).

Houve um terceiro caso de adaptação da informação extraída do MIGRANTE.PT com o *lempos* “vacinas-n”. Com 5 exemplos de ocorrência, 4 dos exemplos fazem uso do *lempos* como “boletim de vacinas” e apenas 1 é o uso do lema “vacina”. Devido à maioria dos exemplos da concordância, o *lempos* foi utilizado como “boletim de vacinas”. Por fim, este fenómeno aconteceu também com a expressão “pedido de reagrupamento”, adotado como “pedido de reagrupamento familiar”. Ambos os *lempos* estão nas listas do anexo 1.

- Expressões multipalavra também foram identificadas como incompletas, depois de consultar as suas ocorrências nas concordâncias. Um exemplo de um caso para estas expressões multipalavra incompletas é a expressão “concessão de autorização”. Ao

olhar para as concordâncias, na realidade, esta é “concessão de autorização de residência”. Outro exemplo da situação descrita seria as expressões multipalavra “exercício de atividade”; “exercício de atividade profissional” e “exercício de atividade profissional independente”. Todas as expressões são, na realidade, a mesma expressão comprovada pela consulta às concordâncias.

Neste caso é possível perceber que o Sketch Engine detetou segmentos diferentes da mesma expressão como expressões multipalavra. Ao verificar estas expressões nas concordâncias é possível apurar que todas as unidades pertencem à mesma expressão.

O processamento automático do Sketch Engine foi testado neste trabalho. As listas foram analisadas e notou-se os diferentes usos dos *lempos* simples e o corte de expressões multipalavra. Estas observações comprovam que, para além do processamento automático, é inevitável uma intervenção manual usando o conhecimento implícito e explícito da língua.

Usando o candidato “exercício” como um exemplo de estudo de um lema utilizado em dois contextos diferentes, realizou-se uma pesquisa do lema na funcionalidade Word Sketch nos dois *corpora*, para se perceber a distribuição e uso da unidade em questão em cada *corpus*. O resultado consistiu num aglomerado de contextos onde o lema “exercício” aparece.

Os gráficos mostram os contextos em que o lema ocorre como uma rede das unidades lexicais mais frequentes junto ao lema “exercício”.

Nota-se em ambos os gráficos as diferenças semânticas entre os *corpora*. Tanto o *corpus* MIGRANTE.PT como o *corpus* MATERIAIS_PLE_2023 têm o lema em questão precedido de verbo ou adjetivo e seguido de verbo ou adjetivo. A diferença consta no seguinte: Os verbos que antecedem o lema são completamente distintos entre os *corpora*. A diferença dos verbos mostra um contexto muito diferente para o lema.

exercício

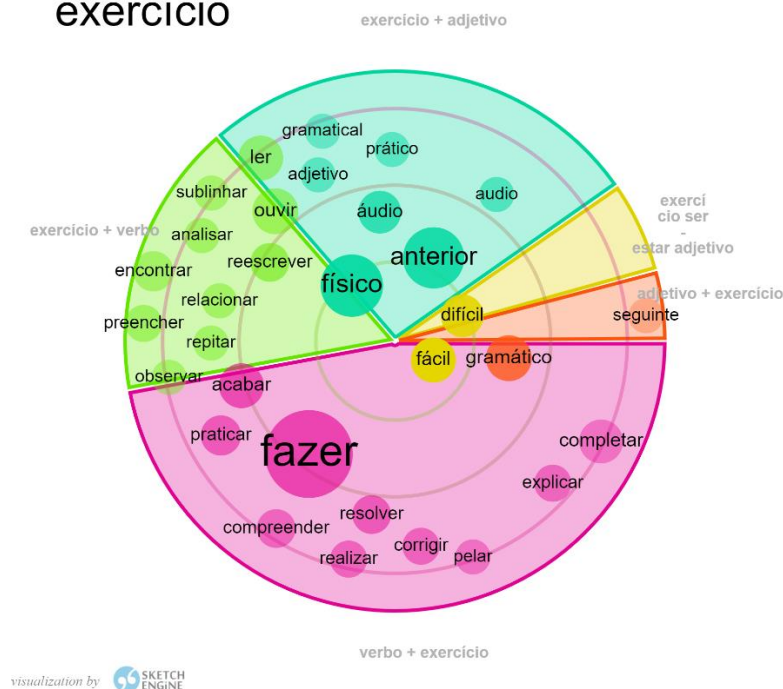


Figura 3: Word Sketch do lema “exercício” no corpus MATERIAIS_PLE_2023.

Os verbos no *corpus* MATERIAIS_PLE_2023 remetem semanticamente para um contexto de sala de aula: “fazer”; “acabar”; “praticar”; “compreender”; “resolver”; “realizar”; corrigir¹⁰; “explicar” e “complementar” são verbos relacionados com a resolução e compreensão de exercícios de ensino/aprendizagem língua estrangeira. O mesmo pode ser dito dos verbos que seguem o lema: “ouvir”; “reescrever”; “relacionar”¹¹; “ler”; “sublinhar”; “analisar”; “encontrar”; “preencher”; “observar”; aplicam-se ao ensino de língua estrangeira.

Os adjetivos também mostram o lema “exercício” a ser utilizado num contexto do ensino do português. Os adjetivos “gramático” e “seguinte” utilizados antes de “exercício”, os adjetivos “físico”; “anterior”; “áudio”¹²; “adjetivo”; “prático” e “gramatical” usados após o lema “exercício” tornam claro a diferença de utilização.

¹⁰ “pelar” é um erro da ferramenta Sketch Engine. Advém da expressão “pelo menos”.

¹¹ “repitar” é uma forma incorreta. A forma em questão é “repetir”.

¹² “áudio” e “áudio” são usadas na mesma forma.

O valor semântico dos verbos e dos adjetivos à volta do lema permitem identificar a unidade lexical de “exercício” em uso no *corpus* de referência, mostrando especificamente com o tema de ensino do português como língua estrangeira.

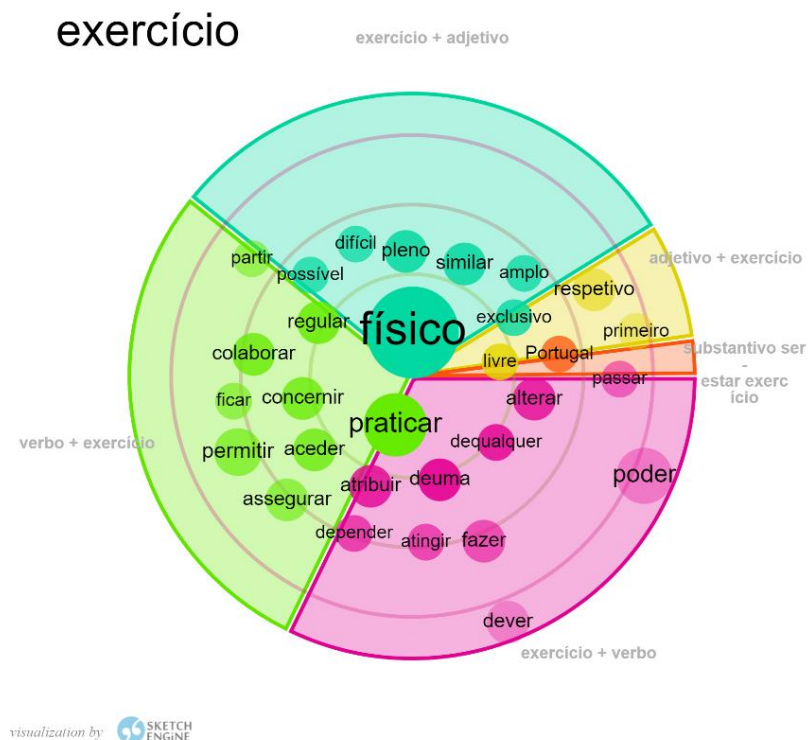


Figura 4: Word Sketch do lema “exercício” no corpus MIGRANTE.PT.

Ao observar a ocorrência de “exercício” no *corpus* MIGRANTE.PT, nota-se o uso de verbos distintos: “praticar”; “regular”; “concernir”; “aceder”; “colaborar”; “ficar”; “permitir”; “assegurar” e “partir”. Estes verbos não invocam o mesmo contexto que no *corpus* de referência, estes verbos apontam para outro campo semântico.

- Um segundo exemplo da presença de unidades idênticas com contextos semânticos diferentes é “mão”. Este lema aparece na lista de expressões multipalavra dentro da expressão “mão de obra”. O lema em questão também é comum no ensino de língua. O domínio da anatomia no ensino de português como língua estrangeira é um dos primeiros grupos temáticos das aulas de língua estrangeira.

Foi comprovado que o lema “mão” aparece em ambos os *corpora*, mas em contextos distintos . Os gráficos do word Sketch mostram o que rodeia o lema nos dois *corpora*, providenciando uma perspectiva aos contextos do lema:

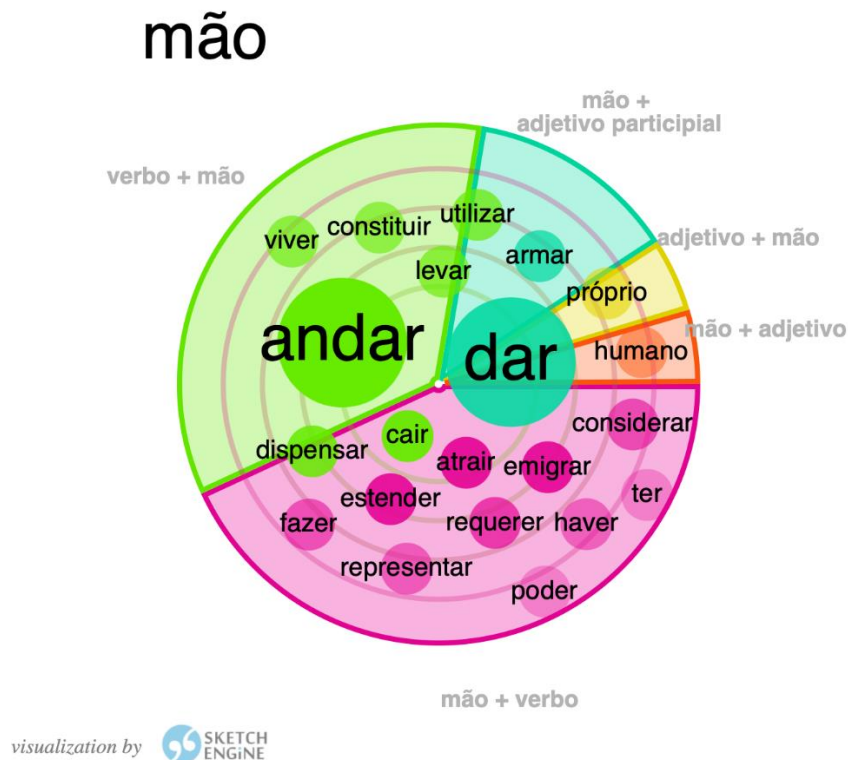


Figura 5: WordSketch do lema “mão” no corpus MIGRANTE.PT.

No *corpus* MIGRANTE.PT, “mão” tem uma rede lexical menor devido à frequência do lema ser menor no *corpus* de foco. A rede lexical é composta por lemas como “armar” de “mão armada”, “humano” de “mão humana”, e o resto da rede contextual apresentada na figura 5 são usos de “mão” dentro da expressão “mão de obra”. Ao consultar a concordância da rede contextual nota-se que muitas das referências a “mão de obra” falam em “mão de obra” imigrante ou emigrante. Isto incorpora o lema dentro do tema migração do *corpus* de foco.

- As listas de *keywords* do *corpus* MIGRANTE.PT mostra unidades lexicais relacionadas com o processo de acolhimento, com uma vertente legislativa, jurídica, informativa e burocrática. A presença da grande maioria das unidades lexicais com este carácter não foi vista nos materiais de ensino do português como língua estrangeira usados no *corpus* MATERIAIS_PLE_2023, como expectável.

Após uma reflexão sobre os materiais de ensino do português como língua de acolhimento disponíveis no capítulo 2.2., abordou-se as tentativas de providenciar o público-alvo migrante com material de ensino de língua. No entanto, a extração de informação do *corpus* MIGRANTE.PT revela expressões relacionadas com a chegada ao país de acolhimento cujo números de frequência apresentam uma representação muito pequena ou uma ausência completa. De acordo com os dados deste estudo, expressões como “proteção internacional”, “agregado familiar”, “pedido de asilo”, “cartão azul ue”, “título de residência”, “encarregado de educação”, estão ausentes do *corpus* MATERIAIS_PLE_2023. Unidades simples como “reinstalação”, “educando”, “imóvel” também aparecem sem representação no *corpus* de referência.

4.3. Organização em campos semânticos

As listas de *keywords* extraídas contém unidades lexicais de várias áreas relacionadas com a realidade das pessoas migrantes em situações de acolhimento.

Para tirar o melhor proveito das listas foi considerada uma organização dos *lempos* extraídos. O objetivo da organização de *lempos* é separar o vocabulário por grupos de unidades lexicais relativos a necessidades de comunicação associadas ao processo de acolhimento.

Em teoria, os campos semânticos podem ser considerados como os grupos de unidades lexicais relacionadas semanticamente (Pereira, 2008). A tarefa aqui é reunir unidades lexicais relacionadas para criar grupos lexicais específicos, que possam servir o propósito de ordenar vocabulário e proporcionar uma consulta mais fácil ao material. A sinonímia, a meronímia, a hiperonímia/hiponímia são relações paradigmáticas entre unidades lexicais, que podem ajudar a indicar a associação semântica das unidades lexicais de um grupo.

Em Hakkı Erten, I., & Tekin, M. (2008) foi feito um estudo para avaliar a retenção de vocabulário apresentado em grupos de palavras sem correlação e em campos semânticos. A conclusão foi que a retenção de vocabulário é comprometida quando os grupos são muito homogêneos. Existe uma retenção menor quando os grupos são feitos com unidades lexicais demasiado parecidas em som ou significado.

Em Nation (2000), trata-se da questão de interferência causada pela apresentação de vocabulário em grupos semânticos. A investigação chegou a uma conclusão semelhante, pois também foi notado uma maior retenção de unidades lexicais em grupos de palavras mais heterogêneos. Nation tirou esta conclusão através de uma observação da literatura existente sobre os campos semânticos, focando-se na interferência de vocabulário utilizando campos semânticos.

Para criar campos semânticos com o vocabulário extraído é preciso tomar em consideração a escolha das unidades lexicais. Contemplando os autores Hakkı Erten, I., & Tekin, M. (2008) e Nation (2000), o objetivo é agrupar o vocabulário por tema, tomando especial atenção à heterogeneidade dos grupos. Quando se fala em heterogeneidade dos grupos, fala-se no mesmo contexto que os autores anteriores. O objetivo foi juntar várias unidades simples de categorias gramaticais distintas e expressões multipalavra. Unidades conectadas pelo contexto de uso com significados próprios. Sendo um vocabulário específico, terá mais facilidade em ser definido concretamente, já que vocabulário específico não tende a ter muita variação em significado.

4.3.1. Definição dos campos relevantes

Para este projeto, a metodologia para a criação de campos semânticos foi observar as áreas de maior importância para a chegada e para a integração de pessoas migrantes e início de vida dessas pessoas no país de acolhimento.

Para avançar com a criação dos grupos semânticos, foi necessário compreender o léxico extraído através de um ponto de situação sobre o acolhimento em Portugal. Este ponto de situação consistiu em consultar fontes especializadas em acolhimento, asilo e imigração para formar os grupos semânticos. Estas fontes foram o Serviço de

Estrangeiros e Fronteiras (SEF); o Conselho Português para os Refugiados (CPR); o Alto Comissariado para as Migrações (ACM) e o Portal Lisboa Acolhe.

A principal entidade a abranger o asilo e refúgio de pessoas migrantes em Portugal é o SEF. O SEF era um serviço de segurança embutido com a responsabilidade do controlo de fronteiras, da proteção internacional, da regularização da situação das pessoas migrantes, da emissão de documentos e outras funções. Esta entidade trata do processo de legalização da situação de pessoas migrantes em situação de asilo nas fronteiras portuguesas, bem como da imigração. O tratamento de assuntos relacionados com a migração também passa pela emissão de passaportes e documentos para a estadia e vivência no país.

O SEF não lida com a inclusão ou integração de pessoas migrantes nas situações de acolhimento, é uma entidade presente na fase da receção das pessoas migrantes em situação de acolhimento.

O CPR dedica-se à receção, à proteção e ao asilo de migrantes à procura de refúgio no território nacional. O CPR tem uma relação direta com o acolhimento. Esta entidade é focada no refúgio de pessoas migrantes que necessitem de proteção internacional e ajuda na instalação e na integração deste grupo de pessoas.

O ACM tem como uma das suas missões a integração e acolhimento de pessoas migrantes. Este instituto público tem a iniciativa de criar várias estratégias para proporcionar auxílio às pessoas migrantes em situação de acolhimento. O ACM usa o seu *website* como ferramenta de ajuda à imigração e ao acolhimento juntando informação sobre as áreas essenciais da vida no país de chegada. O ACM também fornece guias, recursos e outros tipos de materiais para orientar a pessoa migrante na sua integração. O ACM também ajuda no ensino do português como língua não materna e como língua de acolhimento através de cursos de português já referidos anteriormente.

Por fim, existe o Portal Lisboa Acolhe, gerido pelo ACM e pela câmara municipal de Lisboa. O projeto é um apoio à imigração através da comunicação de direitos, deveres, acolhimento, integração, acesso à educação e ao trabalho e até promove a interculturalidade. Este portal junta recursos diretamente relacionados com o tema de imigração, com o propósito de dar ajuda a este grupo de pessoas a viver em Portugal. O ponto forte do portal Lisboa Acolhe é o seu acesso fácil à informação. O portal Lisboa

Acolhe pode não trabalhar diretamente com o processo de chegada e acolhimento de pessoas migrantes, no entanto não foi ignorado devido às suas escolhas de organização de dados sobre a migração para o público-alvo.

Ao consultar os *websites* das entidades do SEF, ACM, CPR e o Portal Lisboa Acolhe, bem como cartazes informativos de centros locais de apoio a pessoas imigrantes (como cartazes de Centros Locais de Apoio à integração de Imigrantes¹³). Estes recursos direcionados à imigração e ao acolhimento proporcionaram indícios nas áreas específicas relacionadas aos grupos semânticos. Foi notado que todas as páginas consultadas separavam as várias áreas de intervenção de ajuda de maneiras muito parecidas. Muitas das entidades têm em comum as áreas mais importantes para a orientação dos migrantes no país de chegada.

Áreas como emprego, cidadania, educação, saúde e até cultura foram notadas como áreas essenciais de ajuda às pessoas migrantes e como áreas de intervenção de todas as entidades visitadas, com ênfase no vocabulário relacionado com documentação, definição de estatutos e processos de aquisição de estatutos.

4.3.2. Os campos semânticos

Para organizar o léxico selecionado a partir das listas de 1000 unidades simples e 1000 expressões multipalavra extraídas do corpus MIGRANTE.PT, criaram-se sete campos semânticos: acolhimento; integração; família; habitação; educação; emprego e saúde. Foram consideradas como ruído os seguintes elementos: nomes de pessoas ou nomes de entidades geográficas, pontuação e erros da ferramenta. Foram analisados 235 *lempos* para serem distribuídos pelos sete grupos semânticos. Foram analisados 111 *lempos* simples e 124 *lempos* multipalavra. Estes números equivalem a aproximadamente 12% das unidades lexicais recolhidas automaticamente. As siglas foram consideradas *lempos* simples pelo Sketch Engine. Portanto dos 111 *lempos* simples, 86 são unidades simples e 25 são siglas.

¹³Os centros de apoio à integração são uma rede criada de uma parceria entre o ACM e várias câmaras municipais do país.

Acolhimento: é um campo semântico que pretende abordar as necessidades imediatas de chegada das pessoas migrantes em situação de acolhimento. O objetivo é agrupar unidades lexicais simples (nomes, verbos e siglas) e expressões multipalavras relativas a processos legislativos, documentos, estatutos, bem como relacionados com a chegada imediata das pessoas migrantes. O grupo “Acolhimento” contém 74 unidades lexicais: 28 unidades simples; 40 expressões multipalavra e 6 siglas.

Integração: pretende auxiliar na residência ininterrompida dentro do país de acolhimento. Tem elementos como os processos administrativos do quotidiano, a procura da cidadania, os serviços, os direitos e deveres da pessoa migrante. Este grupo também inclui siglas, expressões e unidades simples à volta dos serviços sociais, documentos e siglas relacionadas com a vivência no país de acolhimento. “Integração” contém 66 unidades lexicais: 34 unidades simples; 25 expressões multipalavra e 7 siglas.

Família: é um grupo semântico que pretende juntar unidades lexicais relacionadas com o núcleo familiar. Este grupo pretende ajudar famílias em assuntos relacionados com menores e com o reagrupamento familiar. É constituído por expressões de estatutos, de processos legislativos, de unidades simples que se referem a menores e a laços familiares. “Família” é um grupo que contém 16 unidades lexicais: 8 unidades simples e 8 unidades multipalavra.

Habitação: foi criado à volta dos assuntos relacionados com habitação e o processo de arranjar sítio para viver. Contém unidades e expressões sobre arrendamento, compra de imóveis, unidades simples que se referem à habitação. Este grupo contém 11 unidades lexicais: 5 unidades simples; 5 unidades multipalavra e 1 sigla.

Educação: é um campo semântico de unidades lexicais relacionadas com ensino, educação e vocabulário relacionado com processos administrativos. O grupo educação contém 32 unidades lexicais: 7 unidades simples; 21 expressões multipalavra e 4 siglas. É um grupo englobante de léxico sobre o infantário, o ensino básico, o ensino secundário, o ensino superior, equivalências, estatutos e processos administrativos dentro da educação.

Emprego: relaciona-se com o vocabulário envolvido no processo de arranjar trabalho. O emprego é uma área vital para a pessoa migrante, portanto foram separadas unidades

simples e expressões relacionadas com a procura de emprego, o desemprego e o preenchimento de vagas de trabalho. Para este campo semântico há 29 unidades lexicais: 6 unidades simples; 18 expressões multipalavra e 5 siglas.

Saúde: inclui o léxico encontrado no corpus MIGRANTE.PT relacionado com a saúde. Pretende auxiliar em questões de emergências médicas ou as suas situações de saúde. É um grupo que reúne siglas, expressões relacionadas com as infraestruturas de saúde. “Saúde” contém 7 unidades lexicais, 6 expressões multipalavra, 1 sigla e 1 unidade simples. Este campo semântico é o mais pequeno de todos.

Os sete grupos criados a partir das listas de *keywords* extraídas do MIGRANTE.PT foram criados tendo em consideração a sua utilidade para a aprendizagem de língua de acolhimento. Os grupos têm um número variado de unidades simples e expressões multipalavra. Dentro da escolha das unidades simples foi feito um esforço para seleccionar unidades simples de categorias gramaticais diversas, já que esta escolha é facilitada pela função de *lempos* que inclui informação gramatical, para haver uma representação rica de verbos, adjetivos e nomes. Foi possível encontrar vários nomes e alguns verbos, mas a maioria dos adjetivos faziam parte de expressões. As expressões tinham mais utilidade que o adjetivo isolado, o que resultou em escolher as expressões multipalavra em vez do adjetivo.

Todo este trabalho da fase da organização de vocabulário dá seguimento ao uso deste para serem elaboradas entradas na ferramenta Lexonomy para a proposta de glossário e os grupos são aproveitados na macroestrutura do glossário.

4.4. Proposta de glossário

Nos capítulos anteriores foi descrito o processo para obter o vocabulário do português como língua de acolhimento, desde a compilação de *corpora*, à sua comparação, passando pela extração de *keywords* simples e multipalavra e, por fim, à organização de léxico por grupos semânticos.

O último passo na investigação levada a cabo sobre vocabulário para o português como língua de acolhimento é mostrar um recurso lexicográfico para o registo do léxico extraído. Esta aplicação prática consistiu numa proposta de glossário.

A estrutura da proposta de glossário foi feita partir das unidades lexicais obtidas nas fases prévias e tem em conta a organização de vocabulário já apresentada. Das listas de 2000 *lempos* extraídos (simples e expressões multipalavra), cerca de 12% do seu conteúdo foi usado na fase do agrupamento em campos semânticos. A partir dos grupos semânticos elaborou-se o “Português Língua Acolhimento”, uma proposta de glossário feita na ferramenta digital Lexonomy.

A proposta de glossário pretende:

- Ter um registo acessível adequado a iniciantes do português;
- Ter uma interface fácil de usar para quem não tem conhecimento linguístico;
- Ter espaço para a tradução da definição, para o glossário ser mais usável.
- Conter exemplos de textos reais, extraídos do *corpus* MIGRANTE.PT;
- Representar os campos semânticos da entrada;
- Representar as relações semânticas entre palavras.

Para a proposta foram elaboradas 31 entradas lexicográficas com os parâmetros descritos.

4.4.1. Público-alvo

O público-alvo para a proposta de glossário são pessoas migrantes em situação de acolhimento, que necessitem de aprender o português como língua de acolhimento. O recurso também tem em consideração a dificuldade que as pessoas migrantes podem encontrar dentro da posição de chegada a um novo país. As circunstâncias em que estes migrantes se encontram têm um vocabulário muito específico para alguém com poucas capacidades linguísticas de português.

4.4.2. Micro e macroestrutura no Lexonomy

Como microestrutura, as entradas lexicográficas contêm:

- entrada/vedeta
- categoria gramatical
- definição em português, definindo o significado da entrada
- tradução da definição (neste caso para o inglês)
- exemplo real, extraído do *corpus* MIGRANTE.PT

- indicação de sinónimos
- ligação a outras entradas relacionadas.

-

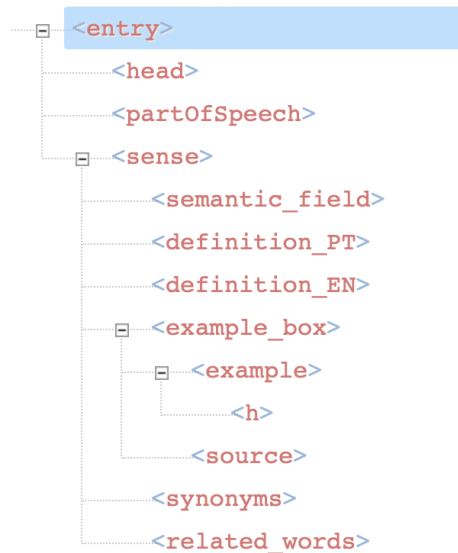


Figura 7: Microestrutura da entrada no Lexonomy.

A Figura 7 mostra a microestrutura das entradas dos Lexonomy. É possível ver os três elementos principais da entrada: <head>, <partOfSpeech> e <sense>. O primeiro elemento <head> é a vedeta, <partOfSpeech> é a categoria gramatical da entrada e <sense> engloba o resto.

Dentro do <sense> é possível ver a indicação do campo semântico (*semantic_field*), o espaço para as definições em diferentes línguas (*definition_PT* e *definition_EN*) e o espaço para exemplos (*example*). Por se tratar de um glossário de um campo específico, não houve entradas com mais de um sentido associado. O que foi definido foi apenas uma definição em português e uma definição em inglês. Os sinónimos (*synonyms*) e palavras relacionadas (*related_words*) estão, naturalmente, associadas ao campo sentido (*sense*).

Em termos de formatação, a microestrutura distingue-se graficamente, como apresentado na Figura 8, abaixo. Nota-se a categoria sintática ao lado da vedeta e a indicação do campo semântico a verde abaixo da vedeta. Nota-se também a representação gráfica dos sinónimos e palavras relacionadas, visivelmente distinta dos outros elementos da entrada.

entrada tester *nome/noun*

Acolhimento

- Esta é uma definição tester para a entrada tester
This is a definition testes for the entry tester.

☞ *Este é o exemplo de **entrada tester**.*

fonte do exemplo

sinónimos|synonyms: recolocação

palavras relacionadas|related words: palavras relacionadas com link

Figura 8: Exemplo da formatação associada à microestrutura concebida

A conexão de entradas relacionadas foi realizada através da criação de ligações diretas como valores do campo <related_words>. Estas funcionalidades são facilitadas pela ferramenta que está a hospedar o recurso lexicográfico.

Para evitar erros de digitação ou variação terminológica, é possível estabelecer listas de valores específicos para alguns campos. Foi o que foi feito para os valores da categoria gramatical, em que se considerou também um valor para expressão (ver Figura 9) e para os valores do campo semântico (ver Figura 10). Considerando o público-alvo, é dada esta informação também em inglês.

Values	
nome noun	noun
verbo verb	verb
adjetivo ad	adjective
Expressão E:	Expression

Figura 9: valores para o campo de categoria gramatical <partOfSpeech>.

Acolhimento	
Integração :	
Emprego Emp.	
Habitação H	
Educação Ed	
Saúde Health	
Família Fam.	

Figura 10: valores para o campo <semantic_field>.

A macroestrutura para a proposta é simples. Uma das limitações do Lexonomy é a falta de personalização na organização das entradas lexicográficas. O Lexonomy apresenta os verbetes por ordem alfabética na página principal e no espaço de consulta das definições e esta ordem é automática. Não é possível ao utilizador fazer uma organização muito personalizada da lista de entradas ou, por exemplo, criar subgrupos de entradas dentro do dicionário digital. Assim, optou-se pela inclusão do campo semântico como elemento da entrada.

O Lexonomy permite, no entanto, fazer uma busca por elementos. É possível definir o que pode ser pesquisado na barra de busca. Como já tinha sido incluído uma etiqueta de campo semântico na entrada (representado por <semantic_field> na microestrutura), foi possível estabelecer o *semantic_field* como elemento pesquisável na barra de busca do Lexonomy. A pesquisa por campo semântico no Lexonomy é uma solução para incluir a organização por campo semântico na ferramenta.

SEARCHING

SEARCHABLE ELEMENTS

The contents of elements you select here will be searchable (in addition to each entry's headword).

- ☐ ENTRY
- ☐ PARTOFSPEECH
- ☐ SENSE
- ☐ H
- ☒ HEAD
- ☐ SOURCE
- ☐ DEFINITION_EN
- ☐ DEFINITION_PT
- ☐ SYNONYMS
- ☐ EXAMPLE_BOX
- ☐ EXAMPLE
- ☐ RELATED_WORDS
- ☒ SEMANTIC_FIELD

Figura 11: Elementos de pesquisa escolhidos no Lexonomy.

4.4.3. Questões relevantes para a compilação

Um dos desafios da compilação do recurso lexicográfico foi como transmitir informação linguística (em particular, o significado) do registo do MIGRANTE.PT de uma maneira clara e concisa a falantes não nativos do português. O objetivo pretendido seria explicar as unidades lexicais extraídas com um registo acessível para alcançar o público-alvo definido.

A linguagem nas definições de sentido foi inspirada no conceito do *Plain English*¹⁴. O recurso precisa de ser compreendido por pessoas migrantes em situações vulneráveis com qualquer nível de conhecimento do português, logo, houve um esforço para criar definições que fossem simples, com um registo um pouco informal, evitando linguagem muito técnica ou específica.

O material trabalhado tende a ser de natureza institucional, por vezes com unidades com significados bastante próximos. Exemplos deste tipo de unidades lexicais são: “imóvel”; “domicílio” e “residência legal”.

Tanto “imóvel” como “domicílio” e “residência legal” são unidades de um nível mais formal que um possível equivalente como “casa”. Embora as formas expressem conceitos próximos (ou equivalentes), a formalidade das unidades tornam-se um

¹⁴ *Plain English*, ou o inglês simples, é um termo que significa a utilização de uma linguagem sem complexidades indicada para qualquer nível de escolaridade ou qualquer nível de conhecimento.

obstáculo a ser ultrapassado devido à dificuldade de apresentação aos falantes não nativos do português.

A relação entre as unidades podem ser de sinonímia (através do elemento <synonyms>) ou uma relação não discriminada, através do elemento palavras relacionadas (<related_words>), que permite facilitar a compreensão de unidades lexicais diferentes, mas semanticamente próximas. O público-alvo pode beneficiar da ligação entre entradas semelhantes, melhorando a aquisição de vocabulário. Mostrar a conexão destas unidades lexicais, defini-las concretamente e mostrar os contextos de utilidade pode ajudar a perceber as diferenças e registo e campo semântico.

A criação dos grupos semânticos pode ajudar na aprendizagem de vocabulário, permitindo perceber melhor o contraste entre unidades próximas. As ligações entre entradas muito próximas em sentido devem ser representadas, pois são benéficas para o ensino, para a compreensão de sinónimos e outras palavras relacionadas.

Fez-se uso da ligação por <related_words> no campo semântico “Habitação” nas entradas referidas anteriormente. Fez-se a ligação de “domicílio” com “imóvel” e “residência legal”.

Considerando o tamanho dos campos semânticos foram usadas 31 das unidades lexicais no Lexonomy. Os sete campos semânticos estão representados no “Português Língua de Acolhimento”. Há 2 entradas do campo “Acolhimento”; 4 entradas de “integração”; 8 de “Habitação”; 3 de “Emprego”; 4 de “Família”; 8 de “Educação” e 2 de “saúde”. Escolheu-se intuitivamente as entradas mais úteis para mostrar a execução da microestrutura e macroestrutura do recurso.

4.4.4. Exemplos de entradas

Considerando as questões levantadas na compilação, e tendo em atenção a utilização de linguagem acessível e clara na construção das definições, o processo de escrever as entradas resultou nas 31 entradas do glossário, incorporadas no Lexonomy. O anexo 3 mostra alguns exemplos do glossário “Português língua de Acolhimento”. Os exemplos foram seleccionados para ilustrar como os cinco tipos de unidades lexicais dos campos semânticos foram adaptados lexicograficamente: nomes; siglas; verbos; adjetivos e expressões multpalavra. Todos os campos semânticos são incluídos no glossário.

“acolhimento”, “salvo-conduto”, “imóvel”, “domicílio”, “creche”, “recrutamento” e “adoção” são os exemplos para as unidades simples. Cada entrada mostra o esforço para criar definições claras para uma melhor acessibilidade. As entradas “imóvel”, “domicílio” e “creche” usam o elemento <related_words> para ligar as unidades do mesmo tema, mas usos diferentes. “domicílio” tem ainda uma representação do elemento <synonyms>. “recrutamento” também utiliza <synonyms> para dar informação adicional sobre o uso real da unidade.

residência legal *Expressão|Expression*

— **Habitação|Housing**
 A morada usada para registo nos serviço de finanças, segurança social e outras instituições.
 The address used for registering at the tax office, social security and other institutions.

ⓘ *Os imigrantes com **residência legal** em Portugal podem trazer um ou mais membros da sua família para o país.*

MIGRANTE.PT

palavras relacionadas|related words: domicílio

Figura 12: entrada de 'residência legal' do glossário Português Língua de Acolhimento.

“NIF” e “IEFP” são entradas de siglas que demonstram como estas foram incluídas no glossário. A definição inclui primeiro o significado da sigla por extenso, depois explica o que é a sigla. No caso de “NIF” explica também o contexto de uso da sigla. Estas entradas também mostram o uso do elemento <synonyms>.

NIF *nome|noun*

— **Integração|Integration**
 Sigla para Número de Identificação Fiscal. O número de identificação fiscal é necessário para identificar uma pessoa nos serviços de finanças, nos serviços de segurança social, nos supermercados e em muitas outras ocasiões.
 Abbreviation for Número de Identificação Fiscal. The Fiscal Identity Number is necessary to identify a person in fiscal services, social security services, supermarkets and many other occasions.

ⓘ *Se não tiver número de identificação fiscal (**NIF**), deve solicitá-lo junto de um Serviço de Finanças.*

MIGRANTE.PT <https://www.acm.gov.pt/pt/-/vou-arrendar-casa-o-que-fazer->

sinónimos|synonyms: número de contribuinte

Figura 13: entrada de 'NIF' do glossário Português Língua de Acolhimento.

“residência legal”, “cartão do cidadão”, “tratamento médico”, “declaração médica”, “jardim de infância” e “união de facto” exemplificam a representação das expressões multipalavra. Em todas as entradas referidas houve o desafio de adaptar as expressões

para o registo pretendido. Os resultados foram definições curtas e claras numa tentativa de passar informação ao público-alvo.

“coabitar” uma entrada que serve para representar o tratamento os verbos nas listas de campos semânticos. Nota-se na definição a tentativa de fazer o significado o mais acessível possível, simplificando o verbo para o público-alvo.

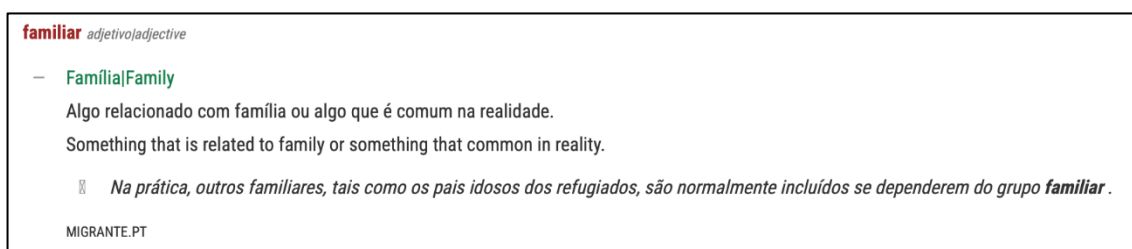


Figura 14: entrada de 'familiar' do glossário Português Língua de Acolhimento.

“familiar” é o exemplo da representação de um adjetivo no glossário. A entrada para o adjetivo é um caso semelhante ao verbo, onde o objetivo de simplificar a linguagem resulta numa definição curta e direta.

5. Considerações finais

O estudo do vocabulário fundamental do português como língua de acolhimento permite ajudar na integração de pessoas migrantes numa situação vulnerável.

Durante a investigação a este vocabulário exerceu-se um conjunto de atividades para obter os resultados preliminares, expostos na metodologia e a seguir nos resultados.

Os resultados da investigação realizada, no contexto do projeto EXPRESSIMI, foram os seguintes:

- A atualização do *corpus* MATERIAIS_PLE_2023, o que permitiu a exploração do *corpus* MIGRANTE.PT (corpus final com 1 105 917 *tokens*).
- A extração de dois tipos de listas (*lempos* simples e expressões multipalavra) com as unidades mais comuns ao MIGRANTE.PT.
- Sete listas de campos semânticos criadas com unidades lexicais do *corpus* de foco.
- A criação da proposta de glossário “Português Língua de Acolhimento”, uma proposta de glossário criado com as unidades lexicais extraídas do MIGRANTE.PT.

Com a atualização do *corpus* de referência é possível chegar à conclusão que os materiais de ensino do português não estão preparados para o ensino a pessoas migrantes em situação de acolhimento e há uma falha no conhecimento linguístico passado a pessoas migrantes. Ao usar o *corpus* MIGRANTE.PT como uma representação do vocabulário usado em situação de acolhimento, comprovou-se a ausência das unidades lexicais relacionadas com situações de refúgio e de asilo, bem como um número limitado de vocabulário à volta da integração e de cidadania. O *corpus* de referência capta uma percentagem da realidade dos migrantes em situação vulnerável. A segunda conclusão tirada sobre o trabalho de *corpora* vem da análise das listas de *keywords*. A linguística de *corpus* traz resultados em termos de frequência, mas ainda requer uma observação manual, conhecimento implícito da língua e perceção da área de especialidade para conseguir informação linguística de qualidade.

Os sete campos semânticos elaborados têm dimensões diferentes. Os campos “Acolhimento” e “Integração” são os de maior dimensão, evidentemente pela natureza do *corpus* de foco. “Habitação”, “Família”, “Educação” e “Emprego” têm um tamanho considerável de unidades lexicais, sendo “Habitação” o mais pequeno dos três. “Saúde” é a lista mais pequena, não surgindo nas listas de candidatos vocabulário do domínio da saúde. O tamanho das listas é diretamente relacionado com os conteúdos do *corpus* de foco. No entanto faz-se a reflexão do grupo de vocabulário para a saúde estar pouco representado. O MIGRANTE.PT não é um *corpus* específico da área da saúde, portanto não é o recurso mais adequado para extrair este vocabulário. Mas é um campo semântico importante que merece mais exploração no ensino do português.

Das listas extraídas foi usado apenas 12%, o que indica potencial para criar mais campos semânticos. O seu conteúdo pode ser mais explorado para recolher, por exemplo, unidades lexicais de natureza legal e criar um campo semântico jurídico.

A heterogeneidade dos campos semânticos foi difícil de criar, pois muitas das unidades analisadas (especialmente nos campos semânticos acolhimento e integração) têm parecenças muito grandes.

O glossário “Português Língua de Acolhimento” é o um dos resultados finais de todos os passos deste trabalho. Esta proposta inclui algumas entradas que representam cada campo semântico. A proposta de glossário apresenta o vocabulário fundamental do

português como língua de acolhimento num recurso lexicográfico. Este recurso lexicográfico teve sempre em consideração o auxílio às pessoas migrantes, numa tentativa de juntar o vocabulário fundamental mais adequadamente para situações vulneráveis. A proposta de glossário finaliza o objetivo do trabalho desenvolvido no projeto EXPRESSI, que era mostrar um recurso de português como língua de acolhimento, feito com unidades lexicais extraídas de *corpora* de língua específicos.

Um exemplo para atividades futuras dentro do português como língua de acolhimento seria ouvir a opinião de formadores e alunos que frequentam os cursos de português como língua de acolhimento, possivelmente na forma de entrevistas ou questionários, para poder aferir dois tipos de informação: se os alunos têm acesso ao vocabulário que precisam; se há vocabulário linguístico está em falta nas aulas. Se sim, qual e como é que o vocabulário extraído neste projeto pode ajudar nas lacunas que possam ser sentidas pelos alunos. Esta informação seria uma mais-valia no futuro do ensino do português como língua de acolhimento.

6. Referências

- Amaro, R., Correia S., Gonçalves M., Barbero C. & Magalhães M. (2022). Inclusive Host Language Teaching: Official Texts for Migrant and Refugee People. In: Meletiadou, E. (ed.) *Policies and Practices for Assessing Inclusive Teaching and Learning*, pp. 183-210. United Kingdom: IGI Global ISBN13: 9781799885795. DOI: <https://doi.org/10.4018/978-1-7998-8579-5.ch009>
- Amaro, R., S. Correia & M. Gonçalves (2021). MIGRANTE.PT. Data set/Database, CLUNL-NOVA FCSH. <https://doi.org/10.34619/ctho-dqgk>
- Bacelar do Nascimento, M. F., P. Rivenc, M.L. Segura da Cruz (1984). *Português Fundamental, Vocabulário e Gramática, tomo 1, vocabulário*. Lisboa: INIC.
- Cotos, E. (2017). Language for specific purposes and corpus-based pedagogy. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning*, pp. 248–264. John Wiley & Sons, Inc. doi:10.1002/9781118914069.ch17
- Elsod, A. & Marques, M. (Coord.). (2019). *Ask the People: a consultation of migrants and refugees*. EMAB, Open Society Foundations.
- Grosso, M. J. (2010). Língua de acolhimento, língua de integração. *Revista Horizontes de linguística aplicada*, 9(2), pp. 61-77.
- Hakki Erten, I., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets. *System*, 36(3), pp. 407-422.
- Hanks, P. (2003). Lexicography. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*, pp.48-69. Oxford: Oxford University Press
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7-36.
- Klein, W. (2001). Lexicology and lexicography. In *International Encyclopedia of the Social & Behavioral Sciences (2nd ed.)*, pp. 938-942. Elsevier.

L'Homme, M.-C. (2014). Why lexical semantics is important for e-lexicography and why it is equally important to hide its formal representations from users of dictionaries. In: *International Journal of Lexicography*, 27(4), pp. 360-377.

Lino, T. (2018). Portuguese lexicography in the internet era. In: Fuertes-Oliveira, P. (ed.), *The Routledge Handbook of Lexicography*, pp. 608-618. Abingdon: Routledge.

Měchura, M. B. (2017, September). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pp. 19-21.

Nation, I.S.P. (2000). *Learning vocabulary in lexical sets: dangers and guidelines*.

Pereira, S. (2008). A importância dos campos léxicos no ensino de língua portuguesa. *InterteXto*, 1(01), pp. 186-208. Brasil: Universidade Federal do Triângulo Mineiro

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tognini-Bonelli, T. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tomé Cornejo, C. (2015). *Léxico disponible. Procesamiento y aplicación a la enseñanza de ELE*. PhD dissertation. Universidad de Salamanca (https://redined.educacion.gob.es/xmlui/bitstream/handle/11162/187841/DLE_Tom%C3%A9Cornejo.pdf?sequence=1).

Lista de Figuras

Figura 1: Composição do corpus MATERIAIS_PLE_2023 por nível de proficiência.....	21
Figura 2: Número de tokens em cada nível de proficiência no corpus MATERIAIS_PLE_2023.	21
Figura 3: Word Sketch do lema “exercício” no corpus MATERIAIS_PLE_2023.	26
Figura 4: Word Sketch do lema “exercício” no corpus MIGRANTE.PT.....	27
Figura 5: WordSketch do lema “mão” no corpus MIGRANTE.PT.....	28
Figura 6: Word Sketch do lema “mão” no corpus MATERIAIS_PLE_2023.	29
Figura 7: Microestrutura da entrada no Lexonomy.	37
Figura 8: Exemplo da formatação associada à microestrutura concebida	38
Figura 9: valores para o campo de categoria gramatical <partOfSpeech>.	38
Figura 10: valores para o campo <semantic_field>.....	39
Figura 11: Elementos de pesquisa escolhidos no Lexonomy.	40
Figura 12: entrada de 'residência legal' do glossário Português Língua de Acolhimento.	42
Figura 13: entrada de 'NIF' do glossário Português Língua de Acolhimento.	42
Figura 14: entrada de 'familiar' do glossário Português Língua de Acolhimento.	43

Anexo 1: Listas de unidades por campo semântico

ACOLHIMENTO	
abrigo (nome)	beneficiário (nome)
abuso (nome)	beneficiário de proteção internacional
ACNUR (sigla)	beneficiário do estatuto de refugiado
acolher (verbo)	Centro de acolhimento
acolhimento (nome)	centro de instalação temporária
apátrida (nome)	CLAIM (sigla)
apatridia (nome)	consulado (nome)
asilo (nome)	CPR (sigla)
asilo político	concessão de asilo
autorização de residência por razões humanitárias	concessão do estatuto de refugiado
autorização de residência provisória	deslocação forçada
autorização de residência temporária	direito de asilo
bem pessoal	direito de proteção internacional
direito ao reagrupamento	repatriamento (nome)
documentação (nome)	requerer (verbo)
documento de viagem	requerente (nome)
entrada em território nacional	requerente de asilo
estada temporária	requerente de proteção internacional
estatuto (nome)	residência (nome)
estatuto de proteção internacional	residência provisória
estatuto de refugiado	residência temporária
extradição (nome)	retirada do estatuto de refugiado
UNICEF (sigla)	salvo-conduto (nome)
visto (nome)	SEF (sigla)
refugiado (nome)	situação irregular
refugiar (verbo)	título de residência
refúgio (nome)	título de viagem
reinstalação (nome)	título de viagem para refugiados
reinstalar (verbo)	visto de curta duração
relocalização (nome)	visto de estada temporária

INTEGRAÇÃO	
acesso (nome)	fasear (verbo)
apoio (nome)	finanças (nome)
apoio social	formalizar (verbo)
aquisição (nome)	habilitar (verbo)
aquisição da nacionalidade portuguesa	identificação civil
autenticação (nome)	indeferir (verbo)
cartão azul UE	inscrição (nome)

Cartão do cidadão	insuficiência económica
certidão (nome)	integração (nome)
cidadão (nome)	IRS (sigla)
comprovativo (nome)	ISV (sigla)
concessão (nome)	IUC (sigla)
conservatória (nome)	IVA (sigla)
dia útil	meio de subsistência
Discriminatório (adjetivo)	montante (nome)
documento de identificação	NISS (sigla)
documento original	notificar (verbo)
emolumento (nome)	pagamento faseado
entidade competente	pedido de reagrupamento familiar
estatuto de residente de longa duração	
pedido de renovação de autorização de residência	requisito (nome)
PLNM (sigla)	residência de longa duração
Portal das Finanças	residente (nome)
prazo (nome)	residente de longa duração
preenchimento (nome)	residir (verbo)
prorrogação (nome)	Segurança Social
prorrogação de permanência	Serviço de Finanças
prorrogar (verbo)	subsídio (nome)
reagrupamento familiar	subsistência (nome)
rubricar (verbo)	titular de autorização de residência
registo (nome)	validação (nome)
rendimento (nome)	valor mensal
requerimento (nome)	veículo (nome)

FAMÍLIA
adoção (nome)
agregado familiar
apadrinhar (verbo)
ascendente (nome)
comparticipação familiar
cônjuge (nome)
emancipar (verbo)
familiar (adjetivo)
laço familiar
membro da família
menor (nome)
poder paternal
progenitor (nome)

responsabilidade parental
subsídio parental inicial
união de facto

HABITAÇÃO
alojamento (nome)
área de residência
arrendamento (nome)
bem imóvel
coabitar (verbo)
domicílio (nome)
IMI (sigla)
imóvel (nome)
renda de casa
residência habitual
residência legal

EDUCAÇÃO	
AEC (sigla)	equivalência (nome)
agrupamento de escolas	estabelecimento de educação
berçário (nome)	estabelecimento de ensino
CET (sigla)	grau académico
ciclo de estudos (sigla)	grau de licenciado
CQEP (sigla)	grau de mestre
creche (nome)	jardim de infância
diploma (nome)	matrícula (nome)
educador de infância	nível secundário
estudante do ensino superior	pedido de equivalência
educando (nome)	prosseguimento de estudos
EFA (sigla)	regulamento interno
encarregado de educação	renovação de inscrição
ensino básico	ingresso (nome)
ensino público	instituição de ensino superior
ensino superior	instituição universitária

EMPREGO
ACT (sigla)
atividade profissional
atividade profissional independente
contrato (nome)
contrato de prestação de serviços

contrato de trabalho
crédito de horas
desemprego (nome)
empregador (nome)
empresa de acolhimento
entidade empregadora
entidade empregadoras
estagiário (nome)
exercício de atividade profissional independente
formação profissional
GAQ (sigla)
GIP (sigla)
IEFP (sigla)
mão de obra
prestação de serviços
recrutamento (nome)
rescindir (verbo)
TCO (sigla)
trabalhador dependente
trabalhador independente
trabalhador sazonal
trabalho sazonal
serviço de emprego
subsídio de desemprego

SAÚDE	
anti-alérgico (nome)	doença infecciosa
boletim de vacinas	peessoa com necessidades
declaração médica	prescrição médica
DGS (sigla)	tratamento médico

Anexo 2: Descrição do glossário (Lexonomy)

PORTUGUÊS LÍNGUA DE ACOLHIMENTO ☆

[EDIT](#)[CONFIG](#)[UPLOAD](#)[DOWNLOAD](#)[LINKS](#)[...](#)

DESCRIPTION

O glossário **Português Língua de Acolhimento** tem como objetivo ajudar as pessoas migrantes em situação de acolhimento a melhorar o seu conhecimento de português. O glossário trata expressões usadas na comunicação com vista ao acolhimento e à integração de pessoas migrantes, incluindo os processos legislativos e burocráticos da vida no país de acolhimento.

O glossário pode ser consultado por entrada e por pesquisa por campo semântico, na barra de pesquisa. As expressões consideradas estão organizadas em seis campos semânticos: *Acolhimento*; *Integração*; *Educação*; *Habitação*; *Saúde* e *Família*. As entradas do glossário incluem informação sobre a categoria sintática (em português e inglês); o campo semântico; a definição (em português e em inglês); exemplos de textos autênticos; sinónimos; e ligações com entradas relacionadas.

O glossário Português Língua de Acolhimento é resultado do trabalho de investigação de Cláudia Castro, realizado no mestrado de Ciências da linguagem da NOVA FCSH, especialidade de Terminologia e Gestão de Informação Especializada, no contexto do projeto EXPRIMI – Exploração preliminar de corpus para ensino inclusivo da língua de acolhimento baseado em tarefas (<https://clunl.fcsb.unl.pt/projetos/projetos-curso/exprimi/>). O projeto EXPRIMI visa, entre outros objetivos, contribuir com métodos de identificação, extração e descrição do vocabulário fundamental para a aprendizagem do Português como língua de acolhimento, utilizando ferramentas e procedimentos computacionais. O vocabulário considerado para o glossário foi extraído do corpus MIGRANTE.PT (<https://doi.org/10.34619/ctho-dqgk>), compilado no âmbito do projeto.

The glossary **Portuguese Host Language** aims to help migrant people in host situation to improve their knowledge of Portuguese. The glossary deals with expressions used in communication with a view to welcoming and integrating migrant people, including the legislative and bureaucratic processes of daily life in the host country.

The glossary can be consulted by entry and by searching by semantic field, in the search bar. The expressions are organized into six semantic fields: *Welcoming*; *Integration*; *Education*; *Housing*; *Health* and *Family*. Glossary entries include information about the syntactic category (in Portuguese and English); the semantic field (in Portuguese and English); the definition (in Portuguese and English); examples of authentic texts; synonyms; and links to related entries.

The Portuguese Host Language glossary is the result of research work by Cláudia Castro, carried out under the scope of the master's degree in Language Sciences at NOVA FCSH, specializing in Terminology and Specialized Information Management, in the context of the EXPRIMI project – *Preliminary exploration of corpus for inclusive teaching of task-based host language* (<https://clunl.fcsb.unl.pt/en/projetos/projetos-curso/exprimi/>). The EXPRIMI project aims, among other objectives, to contribute with methods of identification, extraction and description of the fundamental vocabulary for learning Portuguese as a host language, using computational tools and procedures. The vocabulary considered for the glossary was extracted from the MIGRANT.PT corpus (<https://doi.org/10.34619/ctho-dqgk>), prepared within the scope of the project.

Anexo 3: Entradas do glossário Português Língua de Acolhimento

acolhimento *nome/noun*

— **Acolhimento|Welcoming**

Processo de receber migrantes em situações vulneráveis.

The process of receiving migrants in vulnerable situations.

☞ *Dinamizar uma rede de autarquias de excelência no apoio ao **acolhimento** e integração de refugiados reinstalados.*

MIGRANTE.PT

salvo-conduto *Expressão|Expression*

— **Acolhimento|Welcoming**

Documento oficial usado para fazer a travessia entre países.

Official document used to travel through countries.

☞ *A emissão de **salvo-conduto** com a finalidade exclusiva de permitir a saída do País é da competência do diretor nacional do SEF, com faculdade de delegação.*

MIGRANTE.PT

imóvel *nome/noun*

— **Habituação|Housing**

Espaço para viver. Inclui apartamentos e moradias.

Living space. Includes apartments and houses.

☞ *A Licença de Utilização visa atestar a que uso o **imóvel** se destina e que este se encontra apto para o respetivo fim.*

MIGRANTE.PT

palavras relacionadas|related words: domicílio

domicílio *nome/noun*

— **Habituação|Housing**

Espaço para viver. Também referido por casa. Inclui moradias e apartamentos.

Living space. Also referred to as home. Includes houses and apartments.

☞ *O trabalhador deve anexar ao contrato a identificação e **domicílio** da pessoa ou pessoas beneficiárias de pensão em caso de morte resultante de acidente de trabalho ou doença profissional.*

MIGRANTE.PT

sinónimos|synonyms: casa

palavras relacionadas|related words: Imóvel

palavras relacionadas|related words: Residência legal

coabitar *verbo/verb*

— **Habituação|Housing**

Partilhar espaço com outras pessoas. Estas pessoas podem ser da família ou não.

Sharing a house with other people. These people can be family or not.

☞ *Sim, mas para efetuar o pedido terá de ter autorização de residência válida, e o seu familiar terá de ter entrado legalmente em Portugal, depender **coabitar** consigo.*

MIGRANTE.PT

residência legal *Expressão|Expression*

— **Habituação|Housing**

A morada usada para registo nos serviço de finanças, segurança social e outras instituições.

The address used for registering at the tax office, social security and other institutions.

☞ *Os imigrantes com **residência legal** em Portugal podem trazer um ou mais membros da sua família para o país.*

MIGRANTE.PT

palavras relacionadas|related words: domicílio

cartão do cidadão *nome/noun*

— Integração|Integration

Documento de identificação pessoal

Document for personal identification.

¶ *O pedido de matrícula deve ser apresentado via internet, na aplicação informática disponível no Portal dasEscolas www.portaldasescolas.pt, com o recurso à autenticação através do **cartão do cidadão**.*

MIGRANTE.PT

palavras relacionadas|related words: identificação civil

NIF *nome/noun*

— Integração|Integration

Sigla para Número de Identificação Fiscal. O número de identificação fiscal é necessário para identificar uma pessoa nos serviços de finanças, nos serviços de segurança social, nos supermercados e em muitas outras ocasiões.

Abbreviation for Número de Identificação Fiscal. The Fiscal Identity Number is necessary to identify a person in fiscal services, social security services, supermarkets and many other occasions.

¶ *Se não tiver número de identificação fiscal (**NIF**), deve solicitá-lo junto de um Serviço de Finanças.*

MIGRANTE.PT <https://www.acm.gov.pt/pt/-/vou-arrendar-casa-o-que-fazer>

sinónimos|synonyms: número de contribuinte

tratamento médico *Expressão|Expression*

— Saúde|Health

O ato de receber ajuda médica.

The act of receiving medical help.

¶ *Ontem à noite, o governo italiano autorizou o desembarque de mais 8 migrantes a bordo do navio para receberem **tratamento médico**, depois de 27 menores não acompanhados terem sido autorizados a desembarcar.*

MIGRANTE.PT <https://cpr.pt/open-arms-e-importante-que-as-pessoas-desembarquem-o-mais-depressa-possivel/>

palavras relacionadas|related words: declaração médica

declaração médica *Expressão|Expression*

— Saúde|Health

Documento que prova uma doença e/ou um certo estado físico/mental.

Document that proves a disease and/or a certain physical/mental state.

¶ *Em casos específicos de necessidade de DIETA, devido a doença súbita, os pais terão de trazer uma **declaração médica** comprovativa da necessidade da mesma que refira o período em que esta será necessária..*

MIGRANTE.PT

palavras relacionadas|related words: tratamento médico

creche *nome/noun*

— Educação|Education

Espaço educativo para bebés e crianças de colo dos 0 aos 3 anos de idade.

Educational space for babies and toddlers ages 0 to 3 years old.

¶ *A **Creche** é uma resposta social, eficaz na satisfação das necessidades de bem-estar e de desenvolvimento integral das crianças, num clima de segurança afetiva e física.*

MIGRANTE.PT

palavras relacionadas|related words: Jardim de infância

jardim de infância *Expressão|Expression*

— Educação|Education

Espaço educativo para crianças dos 3 aos 6 anos. O jardim de infância vem antes da escola primária.

Educational space for children for ages 3 to 6. Jardim de infância comes before primary school.

¶ *Oferta Educativa O Espaço "A Criança" – Valência de **Jardim de Infância**, destina-se a todas as crianças com idades compreendidas entre os 3 anos e o ingresso no ensino Básico.*

MIGRANTE.PT

palavras relacionadas|related words: creche

IEFP *nome/noun*

— **Emprego|Employment**

sigla para Instituto do Emprego e Formação Profissional. O IEPF é um centro de emprego que ajuda na procura de trabalho.
Abbreviation for Instituto do Emprego e Formação Profissional. The IEPF is an employment center that helps in job seeking.

☞ *Como posso inscrever-me no Serviço de Emprego do Instituto de Emprego e Formação profissional (**IEFP**) ?*

MIGRANTE.PT

sinónimos|synonyms: **centro de emprego**

recrutamento *nome/noun*

— **Emprego|Employment**

Quando uma empresa procura uma pessoa qualificada para preencher uma vaga de trabalho.

When a company is looking for a qualified person to fill a job opening.

☞ *Os GAE têm ao dispor das entidades empregadoras serviços de apoio ao **recrutamento** de cidadãos.*

MIGRANTE.PT

palavras relacionadas|related words: **IEFP**

adoção *nome/noun*

— **Família|Family**

processo legal para ser o pai ou mãe de uma criança.

Legal process to become the mother or father of a child.

☞ *A nacionalidade pode ser adquirida por efeito da vontade, pela **adoção** e por naturalização.*

MIGRANTE.PT

união de facto *Expressão|Expression*

— **Família|Family**

Quando duas pessoas estão juntas e vivem juntas, mas não são casadas.

When two people are together and live together but are not married.

☞ *2 – Podem ser efetuados inquéritos e controlos específicos quando existam indícios fundados de fraude ou de casamento, **união de facto** ou adoção de conveniência, tal como definidos no número anterior.*

MIGRANTE.PT

familiar *adjetivo|adjective*

— **Família|Family**

Algo relacionado com família ou algo que é comum na realidade.

Something that is related to family or something that common in reality.

☞ *Na prática, outros familiares, tais como os pais idosos dos refugiados, são normalmente incluídos se dependerem do grupo **familiar**.*

MIGRANTE.PT