Linha D'Água (Online), São Paulo, v. 33, n. 1, p. 69-104, jan.-abr. 2020

EXPLORAÇÃO DE *CORPORA* PARA EXTRAÇÃO E DESCRIÇÃO DE LÉXICO DE ESPECIALIDADE: PARA UMA METODOLOGIA SÓLIDA E SUSTENTADA

EXPLOITING CORPORA FOR EXTRACTING AND DESCRIBING SPECIALIZED LEXICON: TOWARDS A SOLID AND SUSTAINED METHODOLOGY

*Chiara Barbero**
Universidade NOVA de Lisboa, Lisboa, Portugal
*Raquel Amaro***
Universidade NOVA de Lisboa, Lisboa, Portugal

Resumo: A exploração de *corpora* para a extração de léxico de especialidade é um método consensual e comum na construção de recursos lexicais. No entanto, as metodologias empregadas não são explicitamente discutidas, dificultando a comparação e a determinação de abordagens robustas. Para preencher essa lacuna, neste artigo apresentamos e discutimos uma metodologia detalhada para extração de léxico de especialidade a partir de *corpora*, conjugando abordagens linguísticas e estatísticas. O método proposto prevê tanto o uso de *corpora* de especialidade como de *corpora* monitores e inclui: i) análise de dados de frequência; ii) extração de concordâncias e colocações; iii) extração de informação de ordem textual, permitindo a extração de unidades lexicais atómicas e multipalavra e de relações semânticas relevantes. Desse modo, o objetivo da metodologia é a determinação de listas de potenciais unidades lexicais de especialidade e de informações relevantes para a sua descrição que permitam uma validação final rápida e eficiente, maximizando o valor informacional da interação com os especialistas.

Palavras-chave: extração de léxico de especialidade; metodologia; *corpora*; concordâncias; colocações

Abstract: *The use of corpora for specialized lexicon extraction is a common and consensual method for building lexical resources. However, the methodologies used to achieve this are not openly discussed, rendering the comparison and determination of robust approaches difficult. In order to fill in this gap, in this paper we present and discuss a detailed methodology for extracting specialized lexicon from corpus, combining linguistic and statistical approaches. The proposed method uses specialized and monitor corpora and comprises i) frequency information analyses; ii) concordances and collocations extraction; and iii) textual organization information; accounting for core single and multiword expressions and salient semantic relations extraction. This way, our goal is the determination*

* Doutoranda da Universidade NOVA de Lisboa – UNL e pesquisadora no NOVA CLUNL, Lisboa, Portugal; chiarabarbero@fcsh.unl.pt

** Professora na Universidade NOVA de Lisboa – UNL e pesquisadora no NOVA CLUNL, Lisboa, Portugal; raquelamaro@fcsh.unl.pt

http://dx.doi.org/10.11606/issn.2236-4242.v33i1p69-104

of a solid and accurate list of potential specialized lexical units that will allow for a swifter final validation and for maximizing the informational value of the interaction with the experts.

Keywords: *Specialized Lexicon Extraction; Methodology; Corpora; Concordances; Collocations.*

Introdução

As metodologias e ferramentas de extração automática ou semiautomática de léxico têm vindo a ganhar uma posição cada vez mais relevante no âmbito da análise linguística e terminológica devido às numerosas aplicações, tanto no contexto de trabalho lexicográfico e terminológico (construção de dicionários/glossários/bases de dados terminológicas/léxicos), como no contexto de tarefas de processamento de língua natural (extração de informação, sumarização, tradução automática, etc.).

Nas últimas décadas, tornaram-se acessíveis a um público cada vez mais amplo de utilizadores mais ou menos avançados uma série de ferramentas (e.g. AntConc[1], SketchEngine[2], Hyperbase[3]) com interfaces bastante intuitivas e filtros e/ou funcionalidades estatísticos e linguísticos que visam, entre outras, apoiar e facilitar a complexa tarefa de extração e análise de léxico.

O processo de extração automática de léxico, no entanto, apesar de ser uma prática que se quer amplamente disseminada, parece ter como base informação que, por se tomar como básica, primária e transversal, raramente se encontra descrita de modo detalhado. Ou seja, os passos necessários entre a recolha de uma coleção de textos/dados e a produção de uma lista organizada de unidades lexicais, ou unidades lexicais de especialidade (ULE)/termos ou, pelo menos, de potenciais candidatos, são raramente elencados e explicados, formando uma espécie de *black box tácita*.

A utilização de *corpora* relativamente pequenos, embora altamente especializados, e a consequente análise de quantidades reduzidas de concordâncias ou listas de palavras para identificação de candidatos a termos podem ser legítimas em

[1] Disponível em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 24 de fev de 2020.

[2] Disponível em: https://www.sketchengine.eu/. Acesso em: 24 de fev de 2020.

[3] Disponível em: http://hyperbase.unice.fr/. Acesso em: 24 de fev de 2020.

alguns trabalhos terminológicos. No entanto, a análise manual de dados de *corpora* já nem configura uma situação comumente aceite, como sugerido em León-Araúz e San Martín (2018):

> For a long time, the only accessible way of analyzing corpus information for terminological work consisted in manually reading concordance lines. This is time-consuming and inefficient because for a given term a terminologist can be confronted with thousands of concordance lines, many of which may not carry any useful information for the terminologist (LEÓN-ARAÚZ; SAN MARTÍN, 2018, p. 94)

No caso dos lexicólogos e lexicógrafos, a situação é ainda mais extrema. Trabalham com quantidades de dados tendencialmente consideráveis e abertas, pelo que se torna inexequível considerar a revisão manual de concordâncias e colocações por parte de linguistas ou a validação exaustiva das ocorrências por parte de especialistas da área.

O desenvolvimento e a partilha de metodologias motivadas que proporcionem o máximo de automatização possível sem comprometer a qualidade da análise torna-se, assim, um objetivo de investigação essencial. O contributo que este trabalho pretende dar enquadra-se precisamente nessa perspetiva, propondo uma metodologia desenvolvida no contexto de exploração de *corpora* do português (PT) e do italiano (IT) do domínio da Arte Pública com vista à extração e descrição de léxico de especialidade e apresentando as opções e motivações que sustentam cada etapa.

1 Constituição de *corpora* para extração e descrição de léxico de especialidade

A utilização de *corpus* como base de descrição de fenómenos linguísticos implica necessariamente a descrição dos objetivos de investigação que o *corpus* pretende sustentar, na medida em que há critérios de seleção e tratamento dos dados que dependem das diretrizes teóricas seguidas e dos objetivos a atingir. Nesse contexto, a presente secção visa a explicitação das opções teóricas e metodológicas que condicionam a constituição e tratamento dos *corpora* usados.

1.1 Objetivos de investigação e enquadramento teórico

A metodologia em discussão no presente trabalho tem como base o objetivo de extração de léxico do domínio da Arte Pública e de informações relevantes para a sua descrição, a partir de *corpora*, no âmbito de um projeto de investigação mais lato que visa a descrição e modelização de léxico de especialidade (LE) no quadro de um modelo relacional de organização do léxico, a WordNet, e a análise da relação entre léxico comum (LC) e LE.

Uma *wordnet* – rede de palavras – (Miller *et al.*, 1990; Fellbaum, 1998; Marrafa *et al.*, 2005) é uma base de dados em formato eletrónico, construída segundo modelos de organização do léxico mental (o modelo WordNet), nos quais o significado de cada item lexical deriva das relações que esse mantém com os outros nós da rede. Contrariamente a recursos lexicais mais tradicionais, a unidade nuclear de uma *wordnet* não é a palavra, a que é associada a lista dos significados que pode denotar, mas sim o *synset*, conjunto de palavras sinónimas que denota um dado significado (ou conceito, no modelo). A noção de unidade nesse modelo é assim de cariz fortemente lexicológica – unidade linguística com determinadas propriedades formais e gramaticais que denota um único significado estável.

Para os objetivos de investigação em causa e para enquadrar a metodologia apresentada, importa, além disso, definir também o enquadramento teórico no que respeita à noção de LE e de ULE. A definição wusteriana de *termo* enquanto etiqueta/verbalização do conceito[4] deixa espaço a algumas dúvidas relativamente à aplicação prática e à operacionalidade da mesma e, apesar de esse trabalho não se enquadrar numa perspetiva estritamente terminológica, por uma questão de clareza, serão apresentadas e contrastadas algumas definições de termo e de unidade lexical especializada que consideramos essenciais para estabelecermos a definição de ULE adequada e operacional aos objetivos em prossecução.

Muitos autores têm definido *termo* por oposição a *palavra*, partindo de diferentes perspectivas: Sager (1990), por exemplo, foca-se na noção de referência:

4 "Para los terminólogos, una unidad terminológica consiste en una *palabra* a la cual se le asigna un concepto como su significado" (WÜSTER, 1998, p. 21).

"The items which are characterized by special reference within a discipline are the 'terms' of that discipline, and collectively they form its 'terminology'; those which function in general reference over a variety of sublanguages are simply called 'words', and their totality the 'vocabulary'" (Sager, 1990, p. 19). Cabré (1999), por outro lado, opõe *termo* a *palavra* em termos pragmáticos:

> Pragmatics is the factor that most significantly differentiates terms from words. Pragmatically, terms and words differ with respect to their users, the situations in which they are used, the topics they communicate, and the type of discourse in which they usually occur (CABRÉ, 1999, p. 36).

Outros autores têm privilegiado a definição de *termo* em relação ao contexto em que esse é usado, ou seja, são *termos* as unidades lexicais que ativam e veiculam um significado de especialidade num contexto específico (Perez; Rizzo, 2014; Vu et al., 2008). No seu conjunto, esses termos constituem o vocabulário fundamental (*core vocabulary*) do domínio de especialidade em questão (Heylen; Hertog, 2015, p. 203). Faber (2012) nota ainda que, de modo geral, a maioria das ULE apresenta formas nominais complexas, usadas no âmbito de domínios técnicos ou científicos específicos, que ativam significados específicos, com propriedades sintáticas ou valências combinatórias específicas (Faber, 2012, p. 22).

Do ponto de vista do Processamento das Línguas Naturais (PLN), podemos distinguir os termos dos não-termos baseando-nos em critérios estatísticos e linguísticos, de acordo com os níveis de *unithood* e *termhood*, ou seja, respectivamente, o nível de estabilidade de combinatórias sintagmáticas (i.e. utilizando medidas de informação mútua que calculam quais os elementos que coocorrem preferencialmente juntos do que isolados) e o nível de relevância no domínio de especialidade (i.e. utilizando o grau de frequência de ocorrência relativa comparada com um *corpus* monitor ou critérios relacionados com a distribuição das unidades ao longo dos textos) (Heylen; Hertog, 2015; Pazienza *et al.*, 2005; Vu *et al.*, 2008). Nesse contexto, o que pretendemos extrair e estudar são as ULE, não sendo o contraste distintivo entre termo e ULE relevante. Para os efeitos do presente trabalho, consideraremos, então, que uma ULE é uma unidade lexical, atómica ou multipalavra, com propriedades sintáticas, combinatórias e de significado específicas, usadas por

especialistas de um dado domínio, tipicamente em situação de comunicação entre especialistas desse domínio, que podem ser caracterizadas por critérios estatístico -linguísticos de *unithood* e *termhood*.

Essa definição motiva, assim, a relevância da constituição e utilização de *corpora*, bem como a importância e relevância do recurso a ferramentas de exploração de *corpora*, com funcionalidades estatísticas e linguísticas.

1.2 Constituição e caracterização dos corpora

O CORPORART[5] (Barbero, 2019), enquanto *corpus* de especialidade, é um repositório bilingue PT/IT comparável, que recolhe um conjunto de textos de especialidade do domínio da Arte Pública e que visa refletir o uso efetivo das línguas PT e IT por especialistas no âmbito da área de especialidade considerada.

Por textos de especialidade, entendemos textos produzidos por especialistas tendo em conta um público-alvo tendencialmente de especialistas, mas que podem também dizer respeito a situações comunicativas heterogéneas que visam um público mais ou menos especialista, por exemplo: especialistas em formação, público conhecedor, público de especialistas e não-especialistas, etc. Considerando a produção no domínio, determinada por inquérito a especialistas, foi possível organizar os textos em 4 tipos:

- textos científicos, produzidos por especialistas em contextos académicos ou, de modo geral, em contextos científicos de investigação, como dissertações, contribuições em conferências, revistas especializadas, entre outros;
- textos de divulgação/disseminação (em contexto de publicações dedicadas ao setor), produzidos por especialistas em contextos mais abrangentes, direcionados a um público amplo, constituído por especialistas de diferentes níveis, como catálogos, livros e capítulos de livros;

[5] Disponível em: https://clunl.fcsh.unl.pt/recursos-em-linha/corporart-corpus-comparavel-pt-it-de-especialidade-no-dominio-da-arte-publica/. Acesso em: 24 de fev de 2020.

- textos legais, que visam regular e enquadrar as práticas ligadas à Arte Pública, produzidos por e para especialistas, como leis e regulamentos e editais de concursos;
- textos técnicos, produzidos por e para especialistas, embora não necessariamente do mesmo domínio, como relatórios de atividades, instruções, etc.

A definição dos critérios de seleção dos materiais é um passo crucial que tem implicações diretas na qualidade e na usabilidade do *corpus* e está diretamente dependente dos objetivos de investigação propostos. Um *corpus* é uma coleção de textos digitais autênticos que resulta de um processo de amostragem de forma a ser representativa de uma língua ou de uma determinada variedade (McEnery *et al.*, 2006, p. 4). O CORPORART, por sua vez, é constituído por:

- textos originais, não sendo consideradas traduções a partir de outras línguas[6];
- textos disponíveis em formato eletrónico, sem recurso a digitalização de textos impressos;
- textos de especialidade, i.e. produzidos por especialistas para um público-alvo maioritariamente de especialistas ou semiespecialistas;
- textos contemporâneos, i.e. produzidos num intervalo temporal entre 2000 e 2018.

A tabela abaixo apresenta em termos quantitativos os *corpora* compilados de acordo com essa metodologia, no contexto do projeto já mencionado (Barbero, 2019).

[6] No contexto do presente trabalho, optámos por não incluir textos traduzidos de forma a minimizar os riscos de estarmos perante situações de decalque e de empréstimos, resultado do processo de tradução e não da produção direta do especialista.

Tabela 1: CORPORART-PT/IT descrição quantitativa

TIPO DE TEXTO	PT (n.º *tokens*)	IT (n.º *tokens*)
Científico	2 667 533	884 839
Disseminação	134 709	171 400
Técnico	0	8 482
Legal	18 364	29 920
TOTAL	**2 820 606**	**1 094 641**

Como já discutido em Barbero (2019), por razões de natureza extralinguística, a divergência quantitativa dos dados dos *corpora* CORPORART-PT/IT é significativa e uma avaliação cabal dos *corpora* só será possível na extensão em que esses sirvam os objetivos de investigação em curso (Giouli; Piperidis, 2002; Lavid, 2005), no caso, a extração de léxico de especialidade. No entanto, de acordo com várias medidas de similaridade entre *corpora* (Schäfer; Bildhauer, 2013), a avaliação e a comparabilidade dos *corpora* constituídos podem ser desde já aferida no que respeita a:

i) qualidade técnica intrínseca – são *corpora* cuja constituição seguiu a aplicação metódica dos mesmos critérios de seleção, descrição e compilação, com validação por especialistas da área (Schäfer; Bildhauer, 2013);
ii) medidas que permitem aferir a similaridade entre *corpora*, como nível de representatividade por saturação lexical (Corpas-Pastor; Seghiri, 2009) – o nível de saturação lexical dos *corpora* foi medido através da utilização da ferramenta ReCor[7] (*Representativeness of Corpora*), Figuras 1 e 2 em anexo, estando garantida a representatividade a partir dos 100 documentos e dos 1,75 milhões de *tokens* para o corpus PT e dos 75 documentos e do 1 milhão de *tokens*, para o corpus IT.
iii) medidas que permitem aferir a similaridade entre *corpora*, como análise simples de frequências altas (Kilgarriff, 2001) – a Tabela 2 em anexo demonstra alto grau de sobreposição entre os *corpora*.

[7] O programa ReCor é apresentado em Corpas-Pastor and Seghiri (2009). Disponível em: http://www.lexytrad.es. Acesso em: 06 de mar de 2020.

iv) análise de *palavras-chave* (Kilgarriff, 2012) correspondentes ao léxico específico de cada *corpus*, obtida por comparação com dados de referência, como detalharemos nas secções seguintes.

1.3 Tratamento e normalização para exploração (semi)automática

Após a seleção dos materiais a incluir no *corpus*, foi necessária a compilação dos textos, incluindo a sua normalização e a codificação dos respectivos metadados. Os ficheiros originais, independentemente do formato (pdf, jpg, png, html, etc.), foram convertidos em formato de texto simples (txt), com recurso a reconhecimento ótico de caracteres sempre que necessário.

Para cada texto foram codificados os seguintes metadados: título, autor, data, repositório/editor, tipo de texto, URL e categorização por subdomínio, sendo essas informações arquivadas num ficheiro à parte. Além disso, a cada ficheiro foi atribuído um nome relativamente transparente, para um fácil reconhecimento e manuseamento (e.g. CART-IT-PhD01.txt).

Em termos de normalização e tratamento inicial, os textos foram limpos para eliminar *ruído* desnecessário. Para tal, foram retirados:

(i) elementos pessoais (e.g. agradecimentos, referências a bolsas ou projetos);
(ii) informações editorias (e.g. fichas técnicas ou normas de direitos de autor);
(iii) biografias dos autores;
(iv) secções em outras línguas (e.g. resumos ou citações);
(v) referências bibliográficas;
(vi) imagens;

e foram convertidos:

(vii) notas de rodapé em notas de fim de página;
(viii) textos em duas colunas em textos corridos de uma coluna de forma a evitar interferências na ordem dos parágrafos.

Após esse processo de limpeza, a versão final do txt tem: o corpo do texto, o resumo e as palavras chaves, o índice analítico, as legendas de imagens e tabelas (normalmente muito ricas quanto a léxico de especialidade), as tabelas e os anexos, se relevantes relativamente ao domínio.

Não houve normalização ortográfica dos textos, para correção de gralhas e erros. A justificação dessa escolha deve-se ao facto de termos identificado nos hápaxe[8] a maior percentagem de erros e, sendo esses erros tipicamente pontuais, que geram palavras "inexistentes" e que não se repetem iguais duas vezes, seria necessária uma revisão quase inteiramente manual e difícil de automatizar com sistemas de verificação ortográfica. Em termos quantitativos, a análise demonstra que a percentagem de hápaxes nos *corpora* utilizados, tanto os de especialidade como os monitores, é bastante reduzida: mínimo de 0,12% no corpus itTenTen e máximo de 1,4% no *corpus* CORPORART-IT (cf. BARBERO, 2019).

O passo final do tratamento inicial inclui a integração na ferramenta de exploração de *corpora*, o Sketch Engine, sendo o processamento superficial (tokenização e anotação morfossintática larga) feito com recurso às funcionalidades integrantes da ferramenta. Tendo em conta o par de línguas de trabalho, o Sketch Engine foi selecionado por permitir o tratamento idêntico de ambos os *corpora*, bem como a introdução de um *corpus* monitor para o PT Europeu, como explicado na secção 2.1.

2 Extração de léxico e de outras informações relevantes

Quando se fala em extração automática de terminologia, na literatura são tipicamente referidas três abordagens principais: linguística, estatística ou híbrida (Drouin, 2003; Lang *et al.*, 2018; Morin; Hazem, 2014; Pazienza *et al.*, 2005; Periñán-Pascual; Mestre-Mestre, 2015; Vu *et al.*, 2008 entre outros).

A abordagem linguística envolve tipicamente os seguintes passos: i) etiquetagem morfossintática; ii) aplicação de *stop lists*[9] de forma a remover falsos candidatos; iii) reconhecimento e padronização de potenciais sequências gramaticalmente

8 Palavra com frequência 1.

9 Listas de palavras funcionais e auxiliares (preposições, numerais, conjunções, verbos auxiliares) com frequências muito altas.

admissíveis e expectáveis (e.g. adjetivo + nome; nome + preposição + nome), ou seja, estritamente relacionadas com a distribuição e a noção de gramaticalidade de cada língua; iv) revisão manual, tanto para juntar significados como para preservar a variação de vária ordem (e.g. letra maiúscula, modificação adjetival *vs*. modificação por sintagma preposicional, etc.). Essa abordagem, apesar de prever a aplicação de filtros linguísticos automaticamente, requer e baseia-se em grande medida em processos de definição e revisão manual, o que torna o trabalho muito moroso e sujeito a avaliação subjetiva.

No entanto, tem vindo a ser recentemente explorada a aplicação de padrões léxico-sintáticos, com base em conhecimento de forma sistemática, sob a forma de filtros linguísticos que facilitam e automatizam a análise das concordâncias para a extração tanto de unidades lexicais relevantes como de relações semânticas estáveis que estruturem modelos relacionais (e.g. *Wordnet*, *Framenet*) e /ou ontologias (Amaro, 2014; Cabezas-García; Faber, 2018; Faber, 2012; Gil-Berrozpe *et al.*, 2017; León-Araúz; San Martín, 2018).

A abordagem estatística, por outro lado, propõe-se medir e equacionar as frequências de ocorrência tanto de unidades isoladas como de mais que uma unidade, calculando a probabilidade de essas ocorrerem juntas, independentemente de qualquer tipo de conhecimento ou propriedade linguística explícita. Essa abordagem, ao contrário da linguística, tem certamente um desempenho melhor em termos de cobertura/tempo de trabalho, ainda que dificilmente consiga manter a mesma qualidade em termos de precisão.

A abordagem híbrida, tal como o próprio nome indica, é uma abordagem que se posiciona entre as anteriormente mencionadas, com o intuito de aproveitar o melhor dos dois mundos, tanto no que diz respeito à cobertura como à precisão dos resultados. As decisões efetivamente implicadas na metodologia híbrida e a forma como essa é aplicada em casos de estudo concretos, no entanto, são muitas vezes pouco claras e descritas.

Os objetivos de investigação globais a montante da metodologia, que incluem dar conta da variação lexical e de significação comum e de especialidade em recursos lexicais relacionais, representam um forte contributo para a desambiguação de situações comunicativas entre especialistas e não-especialistas envolvendo

diferentes níveis de especialização e proficiência. Por exemplo, situações de ensino/aprendizagem; trabalho em equipas interdisciplinares tanto de âmbito académico como profissional; interação entre especialista e clientes ou pacientes; interação entre cidadãos e representantes políticos em contextos diversos, entre outras. Para além do rigor científico, o impacto social desses recursos justificaria por si só o envolvimento e a validação por especialistas na definição dos passos concretos da metodologia. Naturalmente, esse envolvimento deverá ser acompanhado de medidas de minimização de desperdício de tempo, de modo a maximizar o rendimento da interação com os especialistas. Ou seja, mais uma vez, a necessidade de tratamento/triagem semiautomática dos dados de acordo com metodologias e critérios estruturados e sustentados é motivada.

Nesse contexto e, assumindo pelas razões acima indicadas que a combinação de metodologias é o caminho a seguir, a metodologia apresentada é híbrida, aplicada ao caso de estudo em curso, tal como detalhada nas subsecções seguintes no que respeita à extração, bem como na secção 3 no que respeita à análise e tratamento dos dados extraídos.

2.1 Extração de dados de frequência com monitorização

Como mencionado na secção anterior, a frequência é um elemento fundamental na análise de *corpora*. Mas, no que diz respeito às línguas de especialidade, a frequência é essencialmente relevante quando comparada com dados relativos à língua comum. Portanto, uma primeira fase de extração semiautomática das unidades lexicais candidatas a integrar o LE da área da Arte Pública é realizada através da comparação das frequências com *corpora* monitores. Isto significa que, para além dos *corpora* de especialidade descritos na secção 1., são também necessários *corpora* de língua comum, de forma a serem utilizados como *corpora* de exclusão para as frequências mais altas. Os *corpora* monitores foram escolhidos de acordo com (i) a cobertura, (variedade de língua), e (ii) a acessibilidade. Nesse sentido escolhemos utilizar:

- para o IT, o itTenTen 2016[10] (JAKUBÍČEK *et al.*, 2013), *web corpus* do italiano contemporâneo composto por 5 864 495 700 *tokens*, já presente no Sketch Engine e de livre acesso;
- para o PT (Europeu), o CORLEX[11], parte do Corpus de Referência do Português Contemporâneo[12] (Mendes *et al.*, 2012). Excluímos a hipótese de utilizar os *corpora* de português do Sketch Engine pois são compostos maioritariamente pela variedade do PT do Brasil.

O princípio de seleção é simples: se uma dada unidade ocorre com elevada frequência no *corpus* de especialidade e não ocorre, ou pelo menos não ocorre com a mesma frequência (relativa) no *corpus* monitor, poderá ser considerada um válido candidato a ser avaliado. Caso ocorra em ambos os *corpora* com frequência relativa igual ou parecida, será excluída por ter um comportamento característico de língua geral. Os passos que propomos para a exploração das frequências são os seguintes:

1. Extração de candidatos a partir dos 200 *types* mais frequentes entre as principais classes morfossintáticas (Nomes, Adjetivos e Verbos) dos *corpora* de especialidade (CORPORART-PT/IT) que NÃO ocorrem entre os 200 *types* mais frequentes dos *corpora* monitores. No *Sketch Engine* isto requer a extração das listas de palavras de forma individual (i) por categoria morfossintática e (ii) por *corpus*, que serão comparadas manualmente num segundo momento (através da ordenação e comparação de listas em Excel, por exemplo). A determinação de um limite específico está relacionada com as outras etapas da análise de frequência, ver passo 2.
2. Extração de unidades com frequências significativamente diferentes entre o *corpus* de especialidade e o *corpus* monitor. São considerados diferentes intervalos, nomeadamente: diferença elevada (>300%); diferença médio-alta

[10] Disponível em: https://www.sketchengine.eu/ittenten-italian-corpus/#toggle-id-3 Acesso em: 12 de dez de 2019.

[11] Disponível em: https://clul.ulisboa.pt/recurso/lexico-multifuncional-computorizado-do-portugues-contemporaneo Acesso em: 12 de dez de 2019.

[12] Disponível em: https://clul.ulisboa.pt/projeto/crpc-corpus-de-referencia-do-portugues-contemporaneo Acesso em: 12 de dez de 2019.

(>100% <300%) e diferença média (>50% <100%). Esses intervalos são analisados de acordo com a dimensão desejada/esperada do léxico final. Ainda assim, diferenças mais altas terão sempre prioridade sobre as mais baixas.

3. Análise de formas com frequências baixas (), devido à incidência de anáfora em língua de especialidade (Castaño *et al.,* 2002; Wang *et al.*, 2011), com exceção do domínio jurídico (cf. Gotti, 2008). Os hápaxes não são considerados nessa fase do processo, devido à percentagem de erros e gralhas (cerca de 50%) (ver secção 1.1.1.). Pelo contrário, as formas com frequência 2 são incluídas, uma vez que a taxa de erros é em média apenas de 10%.

Esses três passos excluem-se mutuamente. Ou seja: se um dado candidato é extraído no primeiro passo, será forçosamente excluído dos passos a seguir. Importa ainda referir que a extração de dados de frequência utiliza formas e lemas atómicos. As análises que dão conta de unidades multipalavra serão descritas nas secções 2.3.1 e 3.2.1.

2.2 **Extração de colocações**

Para os candidatos extraídos como descrito na secção anterior, são extraídas, numa segunda fase, colocações, de acordo com as funcionalidades da ferramenta. A dimensão das janelas de distância (1 a 6 palavras à esquerda e/ou à direita) é condicionada para extração de potenciais unidades multipalavra – janelas menores –, ou para extração de relações léxico-semânticas (ex. antonímia) –, janelas maiores. Quanto aos parâmetros de extração de colocações no Sketch Engine, definimos:

(i) Medidas estatísticas: MI (por ter bom desempenho com frequências altas); T-*score* (por ter bom desempenho com frequências baixas) e o *logDice* (por ter bom desempenho sobretudo em *corpora* grandes).
(ii) Frequência de ocorrência mínima no *corpus*: 5
(iii) Frequência de ocorrência mínima no intervalo de corte: 5 (PT) e 3 (IT)[13]

[13] A diferença nesse parâmetro deve-se ao facto de o *corpus* IT ser menos extenso do que o *corpus* PT.

A secção 3.2 explora em maior pormenor a análise das colocações para a extração de unidades multipalavra e para a extração de relações semânticas, em particular para a extração da relação de antonímia.

2.3 **Extração de concordâncias com monitorização**

Depois de uma primeira aproximação aos dados segundo o princípio de seleção, quer para unidades lexicais individuais quer para combinatórias multipalavra, (secções 2.1 e 2.2.), **é** necessário avaliar candidatos de alta frequência que, apesar de apresentarem frequências semelhantes no *corpora* de especialidade e monitor, poderão eventualmente denotar significados diferentes.

De acordo com vários autores, as línguas de especialidade são "subcódigos" da língua geral que se sobrepõem parcialmente ao código geral (Araúz *et al.*, 2012; Cabré, 1999; Pearson, 1998), sendo expectável que essa sobreposição se reflita também no léxico. Logo, é possível e provável encontrar unidades que integrem ambos os códigos e, portanto, que estejam presentes em ambos os *corpora*. Esse facto motiva o interesse em considerar também as unidades que, apesar de apresentarem frequências semelhantes em ambos os *corpora*, poderão apresentar traços semânticos diferentes, denotando mais do que um significado de acordo com o contexto de uso. E essa diferenciação pode espelhar-se nos argumentos selecionados pelas unidades e, consequentemente, nas formas que coocorrem com elas e que poderemos recuperar através da análise qualitativa das concordâncias.

Será o caso, por exemplo, de unidades como "arte", "espaço", "trabalhar" ou os adjetivos "público" e "social" que, por ocorrerem entre as 200 palavras mais frequentes tanto no CORPORART-PT como no CORLEX (*corpus* monitor), seriam automaticamente excluídas na primeira fase de extração. No entanto, a análise das concordâncias dessas unidades oferece-nos elementos para podermos assumir a existência de significados diferentes em contexto de especialidade e em contexto geral.

De forma a criar um **método objetivo** de trabalho, definimos os seguintes critérios de análise, divididos por categoria morfossintática, restringindo a an**á**lise às 100 unidades mais frequentes dos *corpora* de especialidade:

1) nós de categoria nominal: ordenar as concordâncias **à** direita (+), pela posição 1 ou 2 (ou seja, +1; +2), de forma a contemplar sequências do tipo *nome+adjetivo* e *nome+preposição+nome*.
2) nós de categoria verbal: ordenar as concordâncias **à** esquerda (-), pela posição -1 e -2, e à direita (+1 e +2), de forma a detetar possíveis padrões argumentais.
3) nós de categoria adjetival: ordenar as concordâncias **à** esquerda, -1, e **à** direita, +1, de forma a detetar diferentes estruturas argumentais.

A seguinte tabela apresenta exemplos das três categorias morfossintáticas.

Tabela 3: Exemplos de concordâncias que espelham diferenças de significado

NOMES		**VERBOS**		**ADJETIVOS**	
CORPORART-PT	CORLEX	CORPORART-PT	CORLEX	CORPOART-PT	CORLEX
espaço ~ público ~ urbano ~ de permanência ~ físico ~ da cidade ~ privado ...	espaço ~ verde ~ limitado ~ geográfico ~ aberto ~ fechado ~ exterior ...	trabalhar artista ~ escultor ~ ~ em equipa ~ em atelier ~ na rua ~ paredes ~ a pedra ...	trabalhar ~ em fábrica ~ nº horas motor ~ mulher ~ ~ a terra ~ em empresa ~ no campo ...	público arte ~ obra ~ espaço ~ escultura ~ edifício ~ concurso ~ ...	público opinião ~ setor ~ administração ~ Ministério ~ saúde ~ relações ~ ...

A ordenação e o ajuste sistemáticos das concordâncias considerando a categoria morfossintática da unidade nó e as características das língua-alvo não evitam a análise manual dos dados, mas permitem uma fácil comparação com dados dos *corpora* monitores e, consequentemente, um trabalho mais ágil e o enriquecimento sustentado das listas de potenciais candidatos a ULE, com a vantagem de simultaneamente forneceram pistas relativas ao significado de modo muito imediato.

2.4 Extração de léxico dos títulos

A par do tratamento e dos processos de extração acima descritos, a exploração dos títulos dos textos de *corpora* de especialidade é relevante para a extração de ULE. Os títulos, nos vários tipos de texto típicos de cada domínio, são cada vez mais um tópico de interesse no que respeita ao estudo das línguas de especialidade (Baicchi, 2003; Roy, 2008; Moore, 2010; Soler, 2011; Méndez *et al*. 2014), desde logo porque permitem a identificação do conteúdo dos textos sem a sua leitura (Hoek, 1981). Devido à produção e **à** acessibilidade massiva de conteúdos, a importância dos títulos tem, aliás, crescido, havendo provas de que muitos especialistas confiam muitas vezes apenas na informação dada no título para integrarem as obras como referências nos seus próprios textos, ou para tomarem decisões técnicas (Bérubé *et al*., 2018). A extração e exploração de títulos para extração de palavras-chave, léxico e terminologia, tem sido assim, desde há algum tempo, uma prática disseminada com sucesso (Hu *et al*., 2005; Poulimenou *et al*., 2014), o que motiva a sua consideração na metodologia aqui apresentada.

A extração de léxico dos títulos implica a compilação de um subcorpus de dimensão muito reduzida, mas de grande relevância, a partir do *corpus* de especialidade já constituído. Isto pode ser conseguido através da etiquetagem dos títulos durante o processo de tratamento inicial e normalização do texto em cada ficheiro (mais moroso), ou pela compilação direta (recorte e armazenamento em ficheiro de texto isolado) dos índices dos livros, teses, catálogos, relatórios e linhas iniciais de artigos de revista e textos curtos. Nesses dois últimos tipos de texto – artigo de revista e texto curto – pode ser relevante compilar também os resumos (*abstracts*) e/ou linhas finais da secção introdutória em que são listadas e apresentadas as secções do artigo, de modo a recuperar informação relevante acerca dos títulos de secções e subsecções nesses tipos de texto, tipicamente sem índice.

O *subcorpus* resultante **é**, por sua vez, objeto dos restantes passos da metodologia, com exceção da extração de frequências (secção 2.1), na medida em que, figurando no título por opção do especialista, o léxico extraído é assim imediatamente considerado relevante para o domínio (Goodman *et al*., 2001; Nagano, 2009).

3 Análise (semi)automática de dados

Da extração automática de dados que, como descrito acima, pode muitas vezes estar limitada pelas funcionalidades e ferramentas disponíveis, é possível partir para a definição de metodologias de exploração de *corpora* que passam pela sistematização dos passos de análise dos dados, para extração de informação menos imediata. Nesse contexto e, considerando os objetivos de investigação que moldaram a metodologia que aqui expomos, nesta secção focamo-nos na extração de relações semânticas a partir de concordâncias e de colocações, discutindo sempre que relevante os fatores condicionantes.

3.1 Análise de concordâncias para extração de relações semânticas

Uma vez identificadas as unidades lexicais ou multipalavra que poderão constituir o léxico de especialidade do domínio em questão, como descrito na secção 2, o passo a seguir, tal como normalmente referido na literatura, é a análise das concordâncias. De facto, as concordâncias – segmentos dos textos ou contextos nos quais as unidades lexicais de especialidade ocorrem – oferecem dados interessantes no que respeita não só a propriedades de valência (de subcategorização, de seleção argumental) que por si só espelham propriedades de significação, mas também no que respeita a propriedades léxico-conceptuais (relacionais, *à la Wordnet*) entre as diferentes unidades lexicais (veja-se, por exemplo, os dados na Tabela 3, acima).

No entanto, a análise manual e exaustiva das concordâncias, nomeadamente para as unidades com frequências muito altas, é pouco viável e, muitas vezes, pouco proveitosa. Desse facto decorre que, ao isolarmos de forma sistemática as concordâncias mais ricas, conseguimos ter acesso de forma mais rápida e estruturada à informação lexical e semântica relevante para a descrição das unidades e necessária à modelização dessas unidades em recursos lexicais relacionais. A hipótese aqui a seguir é, então, que "the hypothesis underlying the use of linguistically rich contexts is that the expression of terminological relationships in texts is made through cue words or structures." (Jacquemin; Bourigault, 2003, p. 570).

A padronização de estruturas léxico-sintáticas mais ou menos fixas (Amaro, 2014; Faber, 2012; León-Araúz; San Martín, 2018), reconstruídas a partir da informação distribucional segundo o princípio de que itens com significado semelhante tendem a ocorrer nos mesmos contextos (AMARO, 2014, p. 3002), e o uso das restrições gramaticais e sintáticas próprias de cada língua, permite recuperar de forma sistemática a informação semântica relevante a partir da análise de concordâncias. Por exemplo, os padrões léxico-sintáticos para extração da relação de hiperonímia, classicamente sob a forma "X é um tipo de Y" ou "X é um Y", e da relação de meronímia, como "X é parte de Y" ou "Y é constituído por", são os mais descritos na literatura (León-Araúz *et al.*, 2016). No entanto, para línguas menos exploradas, como o PT e o IT, há ainda a necessidade de estabelecer conjuntos de padrões previamente testados para a extração das diferentes relações semânticas, preferencialmente a serem disponibilizados em bases de dados ou repositórios de outro tipo. A subsecção que se segue descreve em pormenor a definição e utilização de expressões regulares para exprimir padrões léxico-sintáticos que, por sua vez, são potenciais denotadores de relações léxico-semânticas.

3.1.1 **Expressões regulares e padrões léxico-sintáticos**

A exploração de padrões léxico-sintáticos na análise das concordâncias para a extração de relações semânticas e, consequentemente, de informação relevante para a descrição dos significados das unidades lexicais associada às relações semânticas (equivalência: sinonímia; oposição: antonímia; subtipificação/geral-específico: hiperonímia, hiponímia, instanciação; parte/todo: holonímia, meronímia; definição da estrutura do evento: agente, causa/efeito, etc.) é, tal como referido acima, um passo potencialmente relevante. Nesse sentido, foram adaptados e desenhados padrões para o PT e para o IT, de acordo com a metodologia que a seguir se descreve.

O processo de desenvolvimento e aplicação dos padrões léxico-sintáticos pode ser dividido em 5 fases principais, sendo a ordem em que estsas se realizam não forçosamente linear:

1. recolha de padrões já descritos na literatura[14] e adaptação às línguas de trabalho (PT e IT);
2. construção de novos padrões léxico-sintáticos na sequência de um processo de levantamento e identificação de estruturas recorrentes nos textos;
3. conversão dos padrões em expressões regulares, de acordo com as normas estabelecidas pela ferramenta em uso, no caso, o Sketch Engine;
4. teste dos padrões no Sketch Engine de forma a avaliar a produtividade dos mesmos nos *corpora* de especialidade e nos *corpora* monitores;
5. revisão dos padrões de forma a eliminar/ajustar os padrões com resultados altamente ambíguos, eliminar padrões não produtivos e limitar ao máximo a margem de erro.

O exemplo abaixo ilustra esse procedimento.

(1) Relação de hiperonímia – PT: *X é um Y*

a) Padrão léxico-sintático15:

(det) NOMEX (adj) (adj) (sint. prep) VERBO-SER *um/uma* NOMEY (adj) (adj) (sint. prep)

b) Expressão regular Sketch Engine:

[tag="D.*"] [tag="N.*"] [tag="A.*"]? [tag="S.*"]? [tag="N.*"]? [tag="S.*"]? [tag="N.*"]? [tag="A.*"]? [lemma="ser"] [tag="DI.*"] [tag="N.*"] [tag="A.*"]? [tag="S.*"]? [tag="N.*"]? [tag="S.*"]? [tag="N.*"]? [tag="A.*"]?

c) Resultados: 790 correspondências[16] para o CORPORART-PT e 2000 para o CORLEX.

d) Avaliação e problemas: de acordo com o número de correspondências (produtividade) *vs.* taxa de acerto, o desempenho parece ser bom para as unidades testadas. No entanto, é preciso considerar algumas dificuldades que implicam uma revisão manual, nomeadamente no que respeito à distinção de polissemias (ex.: *escultura*

14 Aguilar et al. (2016); Amaro (2014); Faber (2012); Gil-Berrozpe et al. (2017); Khoo e Na (2006); León-Araúz et al. (2016); León-Araúz e San Martín (2018); Sierra et al. (2008).

15 Na expressão regular, os parêntesis '( )' indicam opcionalidade.

16 Resultados não tratados, incluindo resultados duplicados.

pública enquanto *prática*, e *escultura pública* enquanto *objeto físico*) e identificação de sentidos metafóricos.

(2) Relação de hiperonímia – IT: *X e outro/outros Y*
a) Padrão léxico-sintático:
(det) NOMEX (adj) (adj) (sint. prep) *e altro* (det) NOMEY (adj) (adj) (sint. prep)
b) Expressão regular Sketch Engine:
[tag="NOUN"] [tag="ADJ"]? [tag="PRE.*"]? [tag="NOUN"]? [tag="PRE.*"]? [tag="NOUN"]? [tag="ADJ"]? [lemma="e"] [lemma="altro"] [tag="NOUN"] [tag="ADJ"]? [tag="PRE.*"]? [tag="-NOUN"]? [tag="PRE.*"]? [tag="NOUN"]? [tag="ADJ"]?
c) Resultados: 103 correspondências para o CORPORART-IT e 322 939 para o itTenTen.
d) Avaliação e problemas: uma vez que o *corpus* monitor é substancialmente maior, é normal que esse apresente um maior número de resultados. Apesar do desempenho razoável de acordo com a taxa de acerto (70% para as unidades testadas), esse padrão é ambíguo pois, para além de unidades potencialmente hiperónimas, extrai também unidades co-hipónimas.

As relações tipicamente mais exploradas no que diz respeito à extração automática de informação semântica são as relações taxonómicas (hiponímia/hiperonímia) e as relações de todo/parte (holonímia/meronímia). No entanto, no âmbito deste projeto, foram identificadas estruturas e padrões para a extração automática de outras relações relevantes na modelização do léxico e, em particular, de relações já codificadas no quadro da WordNet.PT (Marrafa *et al.*, 2005), tal como co-hipónimos[17], agente/causa, efeito, instrumento e resultado, e relações de categorização.

Os exemplos a seguir mostram uma primeira aproximação de estruturação de padrões para relações não hierárquicas, nomeadamente para a relação de meronímia e a relação de instrumento/resultado.

(3) Relação de instrumento/resultado – PT: *X serve para/é utilizado/usado para Y*

[17] São co-hipónimas as unidades que têm o mesmo hiperónimo direto. Por exemplo *mamífero, peixe, ave* têm como hiperónimo direto *animal*, logo, são co-hipónimos.

a) Padrões léxico-sintáticos:
i) (det) NOMEX (adj) (adj) (sint. prep) que ser utilizado/usado para VERBOy (det|num) (nome)
ii) (det) NOMEX (adj) (adj) (sint. prep) que serve para VERBOy (det|num) (nome)
b) Expressões regular Sketch Engine:
i)[tag="N.*"][tag="A.*"]?[tag="S.*"]?[tag="N.*"]?[tag="S.*"]?[tag="N.*"]?[tag="A.*"]?[lemma="que"]?[lemma="ser"][lemma="utilizar|usar"][lemma="para"][tag="V.*"][tag="D.*|Z.*"]?[tag="N.*"]
ii)[tag="N.*"][tag="A.*"]?[tag="S.*"]?[tag="N.*"]?[tag="S.*"]?[tag="N.*"]?[tag="A.*"]?[lemma="que"]?[lemma="servir"][lemma="para"][tag="V.*"][tag="D.*|Z.*"]?[tag="N.*"]
c) Resultados: CORPORART: 43 correspondências, COLRELX: 69 correspondências
d) Avaliação e problemas: o padrão é menos produtivo (em termos quantitativos) relativamente aos anteriores e apresenta uma taxa de acerto não muito alta (50%). Os falsos positivos extraídos são maioritariamente sentidos metafóricos ou contextos maiores com estruturas complexas que não são abrangidos por essa configuração léxico-sintática.

(4) Relação de instrumento/resultado – PT: *X serve para/é utilizado/usado para Y*
a) Padrões léxico-sintáticos:
i) (det) NOMEX (adj) (adj) (sint. prep) que ser utilizado/usado para VERBOy (det|num) (nome)
ii) (det) NOMEX (adj) (adj) (sint. prep) que serve para VERBOy (det|num) (nome)
b) Expressões regular Sketch Engine:
[tag="N.*"] [tag="A.*"]? [tag="S.*"]? [tag="N.*"]? [tag="S.*"]? [tag="N.*"]? [tag="A.*"]? [lemma="que"]? [lemma="ter|ser"]? [lemma="utilizar|usar|servir"] [lemma="para"] [tag="V.*|N.*"]
c) Resultados: CORPORART: 77 correspondências, COLRELX: 143 correspondências
d) Avaliação e problemas: o padrão é menos produtivo (em termos quantitativos) relativamente aos anteriores e apresenta uma taxa de acerto não muito alta (50%). Os falsos positivos extraídos são maioritariamente sentidos metafóricos ou contextos maiores com estruturas complexas que não são abrangidos por essa configuração léxico-sintática.

De modo geral, as maiores dificuldades encontradas na padronização de estruturas sintáticas que reflitam a explicitação de informação semântica relevante têm a ver com a criatividade e a variação linguística, pois, apesar de serem relativamente limitados em contexto de discurso de especialidade, esses fenómenos tornam mais difícil a tarefa de fixação de padrões. Em particular, esse é o caso no que respeita a: (i) omissões e retomadas anafóricas que obrigam a recuperar referentes em contextos mais alargados; (ii) distinção de polissemias e (iii) reconhecimento de sentidos metafóricos, que implicam avaliação e revisão manual; (iv) reconhecimento e eliminação de resultados e falsos positivos.

Os pares de unidades relacionadas extraídos nesse passo serão validados pelos especialistas como potenciais nós da rede lexical relacional de especialidade a estabelecer *wordnet*, podendo os elementos que estão relacionados com os candidatos a ULE constituir eles próprios ULE a tratar18, o que permite uma forma indireta de extração.

3.2 **Análise de colocações: potencial e limitações**

Considerando a utilização e o funcionamento de ferramentas de exploração de *corpora*[19], há que ter em conta os processos de extração de colocações por elas utilizado por um lado e os resultados apresentados pelas ferramentas por outro. Nesse contexto, a definição de *colocação* tem uma natureza pragmática e funcional: uma colocação é a ocorrência de duas palavras/*tokens* com relevância estatística, considerado um intervalo relativamente curto de proximidade (tipicamente até 5 palavras à direita e à esquerda da palavra nó). No seguimento dessa definição e, como já observado em Sinclair (1991, p. 170), as colocações podem ser dramáticas e interessantes por serem inesperadas, logo, muitas vezes, imprevisíveis ou relevantes para a estrutura lexical da língua por serem frequentemente repetidas. Em qualquer dos casos, têm de ser tidos em conta na análise da estrutura lexical fenómenos muito frequentes em contextos de uso real da língua.

18 É razoável assumir que, em última instância, em nós mais altos da rede, e logo gerais e subespecificados, haverá relação entre ULE (hipónimos) e ULC.

19 Sketch Engine e CPQWeb (Hardie, 2012).

Numa primeira abordagem, para efeitos de extração de léxico de especialidade, o que parece ser mais relevante diz respeito ao facto de as colocações estarem normalmente relacionadas com a colocalização **de palavras e ao facto de ess**as não poderem ser livremente e arbitrariamente combinadas pela utilização das regras da gramática apenas (Pecina, 2009, p. 11). **É por isso** expectável que haja colocações de especialidade ou, por outras palavras, colocações que espelhem ou digam respeito a significados de especialidade. No entanto, os resultados apresentados pelas ferramentas denotam também outro tipo de relações.

A análise das colocações extraídas tendo em conta os objetivos de investigação que motivam a definição da presente metodologia – extração e descrição de léxico de especialidade num modelo relacional – permite-nos, assim, considerar dois tipos distintos de resultados relevantes: identificação de expressões compostas – *multiword expressions* (SAG *et al.*, 2002) (unidades multipalavra) – e extração de relações semânticas pertinentes para a descrição do significado do léxico de especialidade. A presente subsecção apresenta assim a metodologia de análise das colocações extraídas usando ferramentas e métodos automáticos disponíveis, considerando, sempre que possível, as suas limitações e o seu potencial.

3.2.1 Para extração de unidades multipalavra

Tendo em conta as línguas de trabalho, no caso de base, PT e IT, línguas românicas com estratégias de composicionalidade sintagmática semelhantes, considerou-se suficiente o uso de uma janela de – 2 para a esquerda e + 2 para a direita, de forma a abranger estruturas do tipo: Nome + Nome; Nome + Adjetivo(s); Adjetivo + Nome; Nome + Preposição + Nome.

Quanto aos parâmetros de extração de colocações no Sketch Engine, definimos:

(i) Medidas estatísticas: MI (por ter bom desempenho com frequências altas); T-*score* (por ter bom desempenho com frequências baixas) e o *logDice* (por ter bom desempenho sobretudo em *corpora* grandes).

(ii) Frequência de ocorrência mínima no *corpus*: 5

(iii) Frequência de ocorrência mínima no intervalo de corte: 5 (PT) e 3 (IT)[20]

As diferentes categorias morfossintáticas são extraídas de forma isolada. No entanto, uma vez que a maioria das colocações extraídas conjugam diferentes categorias morfossintáticas, poderá haver sobreposições nos resultados.

Os resultados obtidos com a aplicação das medidas estatísticas indicadas são filtrados tendo em consideração as 100 colocações com frequência mais alta, sendo excluídas as colocações com palavras funcionais e gramaticais, de forma a eliminar colocações semanticamente irrelevantes, mas com frequências elevadas (e.g. "do artista", "artista que", "algum artista" etc.). Entre essas, ainda, são eliminadas as combinatórias composicionais que não respeitem as estratégias de composicionalidade das línguas (e.g. "artista e arquiteto"), para chegar a uma lista de candidatos mais limitada a ser submetida à avaliação dos especialistas.

Em relação a essa primeira fase de extração e análise de colocações, a metodologia foca-se em particular na extração de ULE multipalavra, na ótica de otimizar o trabalho e evitar processos morosos de análise manual. A análise de colocações para extração de relações semânticas é descrita na subseção seguinte.

3.2.2 **Para extração de relações semânticas**

A análise de colocações para extração de relações semânticas é um passo extra na incrementação de lista de candidatos a ULE, por um lado, e de extração de informação relevante para a descrição dessas unidades num léxico relacional, por outro. A inclusão desse passo específico de análise prende-se com trabalho realizado sobre a antonímia (Amaro, 2019), na medida em que essa se reveste de um caráter eminentemente colocacional.

A antonímia pode ser definida como compreendendo uma ampla gama de situações de oposição de significado (Justeson; Katz, 1991) ou, por sua vez, pode ser usada para descrever uma conexão rígida entre formas específicas de palavras (Cruse, 2000; Vossen, 2002). No entanto, independentemente dos postulados e suposições

[20] A diferença nesse parâmetro deve-se ao facto de o *corpus* IT ser menos extenso do que o *corpus* PT.

teóricas, é consensual que as colocações são úteis para encontrar antónimos bem estabelecidos e podem ser usadas para categorizá-los (Muehleisen, 1997; Jones, 2002; LEE, 2013). Amaro (2019) defende que a relação de antonímia é uma relação definida por propriedades colocacionais: depende de propriedades de significado específicas (Lyons, 1977; Jackson, 1988), mas exige um alto grau de "atração textual" e frequência combinatória, que não se verifica para outras relações (Amaro, 2014), cumprindo o Princípio Idiomático de Sinclair (Sinclair, 1991). Constitui, assim, um caso ideal de estudo das colocações para extração e descrição de léxico.

Dado que os antónimos são da mesma categoria morfossintática, em termos de metodologia é necessário considerar a informação morfossintática i) na extração de concordâncias e ii) na filtragem de resultados. Daqui resulta o seguinte procedimento:

1. Extrair concordâncias dos itens selecionados por frequência (secção 2.1 e 2.3)
   - usando o lema (e não a forma) para nomes e adjetivos; iterar com formas flexionadas para confirmação;
   - usando forma flexionada (não infinitiva) para verbos.
2. Extrair as colocações a partir das concordâncias, usando uma janela de -3 e +3, numa primeira iteração, e alargando se necessário.
3. Analisar as colocações (até à posição 50), ordenadas por ordem de relevância (de acordo com a medida estatística usada – MI com frequências altas; T-*score* com frequências baixas e o *logDice corpora* grandes) e selecionar como candidato o primeiro resultado da mesma categoria morfossintática com significado potencialmente oposto.

A tabela 4 lista alguns exemplos, extraídos com a mesma metodologia do Corpus de Referência do Português Contemporâneo (CRPC), via CPQWeb[21], e do CORPORART-PT, via Sketch Engine.

[21] Disponível em: http://alfclul.clul.ul.pt/CQPweb/portugal/index.php?thisQ=search&uT=y Acesso em: 12 de dez de 2019.

Tabela 4: Exemplos de antónimos extraídos

	CRPC	Posição na lista	CORPORART-PT	Posição na lista
Nomes	*velho – novo*	**7ª**		
Verbos	*morrer – matar* *– viver*	**5ª** **9ª**	*valorizar – desvalorizar*	**7ª**
Adjetivos	*bonito – feio* *público – privado*	**34ª** **16ª**	*individual – coletivo* *público – Ø* *efémero – permanente*	**1ª** **2ª**

Os exemplos acima demonstram o potencial de análise das colocações para extração de relações semânticas e para descrição do léxico de especialidade *vs*. léxico comum. Por exemplo, relativamente ao adjetivo *público*, é possível verificar que no domínio de especialidade da Arte Pública a coocorrência com o antónimo de LC, "privado" não surge nas primeiras 50 posições, levando-nos à hipótese de que outras colocações (que espelham unidades multipalavra como Arte Pública) serão mais relevantes e que o significado de *público* nesse domínio não é exatamente o mesmo que no domínio geral. Apesar de ainda pouco trabalhado no que respeita aos dados do projeto que enforma a presente metodologia, o resultado inicial da análise de colocações para extração de relações semânticas mostra-se bastante promissor, com resultados de interesse que motivam a sua integração na metodologia.

Conclusões

A determinação e a descrição de léxico de especialidade com garantia de qualidade exigem a validação por especialistas das unidades extraídas e tratadas. Por esse motivo, a metodologia para extração de léxico e de relações léxico-semânticas aqui detalhada visa reduzir ao mínimo o tempo e o esforço despendido no processo de validação, maximizando o valor informacional da interação com os especialistas a partir da exploração e recolha de informação a partir de *corpora*. A Figura 1 sintetiza de modo esquematizado a metodologia definida.

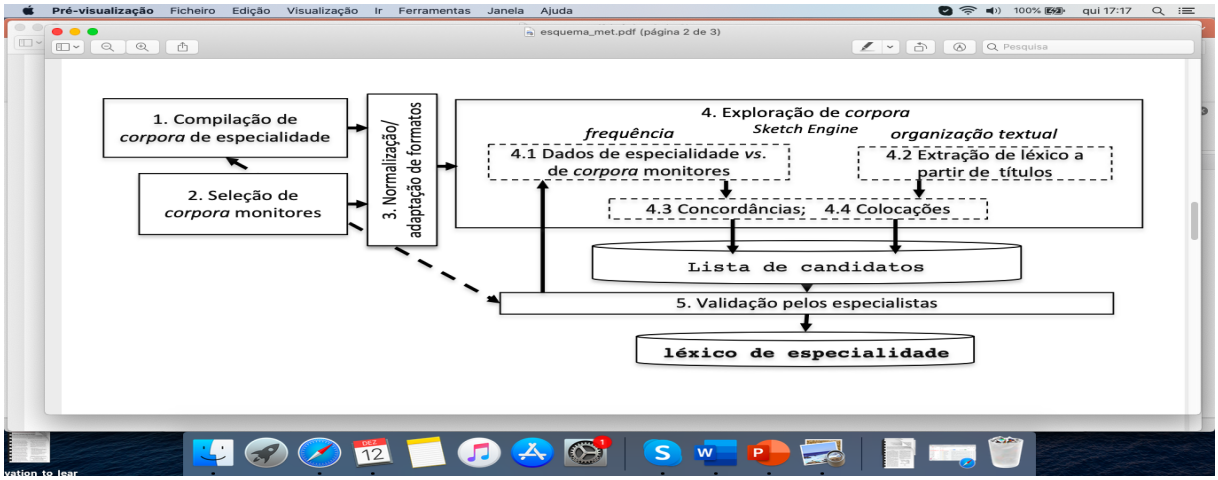

Figura 1: metodologia de extração e descrição de léxico de especialidade

A metodologia apresentada faz uso de várias informações linguísticas – categoria morfossintática, padrões léxico-sintáticos, organização textual – e de ferramentas e métodos de exploração de *corpus* disponíveis – frequências, concordâncias, colocações –, definindo explicitamente as etapas de extração de candidatos a ULE e de extração de relações semânticas relevantes para validação pelos especialistas. Desse modo, contribuímos para o preenchimento de uma lacuna existente na área, promovendo uma discussão aberta e necessária acerca dos métodos para a extração de léxico de especialidade a partir de *corpora*.

Referências

AGUILAR, C.; Acosta, O.; Sierra, G.; Juárez, S.; Infante, T. Extracción de contextos definitorios en el área de biomedicina. *Procesamiento de Lenguaje Natural*, n. 57, p. 167–170, 2016.

Amaro, R. Extracting semantic relations from Portuguese corpora using lexical-syntactic patterns. *Proceedings of the 9th LREC Conference 2014*, p. 3001–3005, 2014.

_____. Antonymy as a collocational relation: analysis and implications for lexicographic resources. *Workshop on collocations: Collocations in Lexicography: existing solutions and future challenges*, eLex 2019, Sintra, Portugal, 2019.

Baicchi, A. Relation complexity of titles and texts: A semiotic taxonomy. In: Merlini Barbaresi, L. (Ed.). *Complexity in Language and Text*. Pisa: PLUS, 2003, p. 319–341.

Barbero, C. CORPORART – um corpus de arte pública para a extração de léxico: representatividade e comparabilidade em corpora de especialidade. *Revista da APL*, n. 5, p. 43–57, 2019.

Benson, B. Collocations and idioms. In: Ilson, R. (Ed.). *Dictionaries, Lexicography and Language Learning.* Oxford: Pergamon, 1985, p. 61-68.

Bérubé, N.; Sainte-Marie, M.; Mongeon, P.; Larivière, V. Words by the tail: Assessing lexical diversity in scholarly titles using frequency-rank distribution tail fits. PLoS ONE, n. 13, v. 7, 2018.

Cabezas-García, M.; Faber, P. Phraseology in specialized resources: an approach to complex nominals. *Lexicography*, n. 5, v. 1, p. 55–83, 2018.

Cabré, M. T. *Terminology: theory, methods and applications.* Amsterdam – Philadelphia: John Benjamins Publishing Company, 1999.

Castaño, J.; Zhang, J.; Pustejovsky, J. Anaphora resolution in biomedical literature. *Symposium on Reference Resolution for Natural Language Processing*, Alicante, Spain, 2002.

Corpas-Pastor, G.; Seghiri, M. Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish). *Corpus use and translating: corpus use for learning to translate and learning corpus use to translate*, n. 82, p. 75-107, 2009.

Cruse, D. *Meaning in Language*. Oxford: Oxford University Press, 2000.

Drouin, P. Term extraction using non-technical corpora as a point of leverage. *Terminology.* International Journal of Theoretical and Applied Issues in Specialized Communication, n. 9, v. 1, p. 99–115, 2003.

Faber, P. *A cognitive linguistics view of terminology and specialized language*. Berlin – Boston: De Gruyter Mouton, 2012.

Fellbaum, C. *Wordnet: an electronical database*. Cambridge: MIT Press, 1998.

Fontenelle, T. What on earth are collocations? *English Today*, n. 4, v. 10, 42-48, 1994.

Gil-Berrozpe, J. C.; León-Araúz, P.; Faber, P. Specifying Hyponymy Subtypes and Knowledge Patterns: A Corpus-based Study. *Proceedings of the 5th eLex Conference*, p. 63–92, 2017.

Giouli, V.; Piperidis, S. Corpora and HLT. Current trends in corpus processing and annotation, *Proceedings of the IJCAI-99*. Insitute for Language and Speech Processing, Bulgaria, p. 1–15, 2002.

Goodman, N.; Thacker, S.; Siegel, P. What's in a title? A descriptive study of article titles in peer reviewed medical journals. *Science Editor*, n. 24, v. 3, 2001.

Gries, S-T. 50-something years of work on collocations. *International Journal of Corpus Linguistics*, n. 18, v. 1, p. 137-166, 2013.

Hardie, A. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, n. 17, v. 3, 380–409, 2012.

Heylen, K.; Hertog, D. Automatic term extraction. In: Kockaert, H. J.; Steurs, F. (Eds.). *Handbook of Terminology*. Philadelphia: John Benjamins Publishing Company, 2015, p. 203-221.

Hoek, L. H. *La marque du titre.* The Hague: Mouton, 1981.

Hu, Y.; Li, H.; Cao, Y.; Teng, L.; Meyerson, D.; Zheng, O. Automatic extraction of titles from general documents using machine learning. *Information Proc. & Management*, n. 42, v. 5, p. 1276-1293, 2005.

Jackson, H. *Words and their Meaning*. London: Longman, 1988.

Jacquemin, C.; Bourigault, D. Term extraction and automatic indexing. In: Mitkov, R. (Ed.). The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press, 2003, p. 568-581.

Jakubíček, M.; Kilgarriff, A.; Kovář, V.; Rychlý, P.; Suchomel, V. The TenTen Corpus Family. *7th International Corpus Linguistics Conference,* 125–127, 2013.

Jones, S. *Antonymy: A corpus-based perspective.* London – New York: Routledge, 2002.

Justeson, J. S.; Katz, S. M. Redefining Antonymy: The Textual Structure of a Semantic Relation. *Literary and Linguistic Computing*, n. 7, p. 176–84, 1991.

Kilgarriff, A., Getting to know your corpus. *Text, Speech and Dialogue – 15th International Conference*, Springer, Heidelberg, p. 3–15, 2012.

Kilgarriff, A. Comparing corpora. *Int. J. corpus Linguist*, n. 6, p. 97–133, 2001.

Kagan, J. *The Nature of the Child*. New York: Basic Books, 1984.

Khoo, C. S. G.; Na, J.-C. Semantic Relations. *Information Science*. Annual Review of Information Science and Technology, n. 40, p. 157–228, 2006.

Lang, C.; Schneider, R.; Suchowolec, K. Extracting Specialized Terminology from Linguistic Corpora. In: FUß, E.; KONOPKA, M.; TRAWINSKI, B.; WAßNER, U. H. (Eds.). *Grammar and Corpora*. Heidelberg: Heidelberg University Publishing, 2018, p. 425–434.

Lavid, J. L. Lenguaje y nuevas tecnologías: nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI. *Epos – Rev. Filol. XX–XXI*, p. 439–445, 2005.

Lee, H-K. Antonymy and gradability: A corpus-based approach on English gradable antonyms. *Linguistic Research,* n. 30, v. 2, p. 335-354, 2013.

León Araúz, P.; Faber, P.; Montero Martínez, S. Specialized language semantics. In: FABER, P. (Ed.). *A cognitive linguistics view of terminology and specialized language.* Berlin: De Gruyter Mouton, 2012, p. 133-211.

León-Araúz, P.; San Martín, A. The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. *Proceedings of the LREC 2018 Workshop Globalex 2018 – Lexicography & WordNets*, p. 94–99, 2018.

León-Araúz, P.; San Martín, A.; & Faber, P. Pattern-based Word Sketches for the Extraction of Semantic Relations. Proceedings of the 5th International Workshop on Computational Terminology, p. 73–82, 2016.

Lyons, J. *Semantics*. 2 vol. Cambridge: Cambridge University Press, 1977.

Marrafa, P.; Amaro, R.; Chaves, R. P.; Lourosa, S.; Martins, C.; Mendes, S. WordNet.PT – Uma rede léxico-conceptual do Português on-line. *XXI Encontro da APL*, 2005.

McEnery, T.; Xiao, R.; Tono, Y. *Corpus-based language studies: An advanced resource book.* New York: Taylor & Francis, 2006.

Mendes, A.; Généreux, M.; Hendrickx, I.; Pereira, L.; Bacelar do Nascimento, M. F.; Antunes, S. CQPWeb: uma nova plataforma de pesquisa para o CRPC. *Textos Selecionados do XXVII Encontro Nacional da APL,* p. 466–477, 2012.

Méndez, D. I.; Alcaraz, M. A.; Salager-Meyer, F. Titles in English-medium Astrophysics research articles. *Scientometrics,* n. 98, v. 3, p. 2331–2351, 2014.

Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K. Introduction to WordNet: An on -line lexical database. *International Journal of Lexicography*, n. 3, v. 4, p. 235–244, 1990.

Moore, A. What's in a title? A two-step approach to optimisation for man and machine. *Bioessays*, n. 32, v. 3, p. 183–184, 2010.

Morin, E., & Hazem, A. Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. *Proceedings of the 52nd Annual Meeting of ACL*, n. 1, p. 1284–1293, 2014.

Muehleisen, V. *Antonymy and Semantic Range in English.* Ph.D. dissertation (Doctor of Philosophy). Northwestern University, Evanston, Illinois, 1997.

Nagano R. L. Lexical comparison of journal article titles in soft disciplines. *Porta Lingua*, p. 111–117, 2009.

Pazienza, M. T.; Pennacchiotti, M.; Zanzotto, F. M. Terminology extraction: An analysis of linguistic and statistical approaches. *Knowledge Mining*, p. 255–279, 2005.

Pearson, J. *Terms in context*. Philadelphia: John Benjamins Publishing Company, 1998.

Pecina, P. Lexical Association Measures: Collocation Extraction. *Studies in Computational and Theoretical Linguistics*, n. 4, 2009.

Perez, M. J. M.; Rizzo, C. R. Assessing four automatic term recognition methods: Are they domain- dependent? *English for Specific Purposes World,* n. 15, v. 42, 2014.

Periñán-Pascual, C.; Mestre-Mestre, E. M. DEXTER: Automatic Extraction of Domain-Specific Glossaries for Language Teaching. *Procedia – Social and Behavioral Sciences*, v. 198, p. 377–385, 2015.

Poulimenou, S.; Poulos, M.; Papavlasopoulos, S.; Stamou, S. Keywords Extraction from Articles' Title for Ontological Purposes. *Proceedings of the International Conference on PMAMCM 2014*, p. 120-125, 2014.

Roy M. Du titre littéraire et de ses effets de lecture. *Protée*, n. 36, v. 3, p. 47–56, 2008.

Sag I. A.; Baldwin T.; Bond F.; Copestake A.; Flickinger D. Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (Ed.). *Computational Linguistics and Intelligent Text Processing.* New Delhi: Springer, 2002, p. 1-15.

Sager, J. C. *A practical course in terminology processing*. Amsterdam – Philadelphia: John Benjamins Publishing Company, 1990.

Salager-Meyer F.; Alcaraz-Ariza, M. A. Titles are "serious stuff": a historical study of academic titles. *JAHR – European Journal of Bioethics,* n. 4, v. 7, p. 257–271, 2013.

Schäfer, R.; Bildhauer, F. *Web Corpus Construction*. Toronto: Morgan & Claypool publishers, 2013.

Seghiri, M. *Compilación de un corpus trilingüe de seguros turísticos* (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad. Tese de doutoramento. Universidad de Málaga, Málaga, 2006.

Sierra, G.; Alarcón, R.; Aguilar, C.; Bach, C. Definitional verbal patterns for semantic relation extraction. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication,* n. 14, v. 1, p. 74–98, 2008.

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

Soler, V. Comparative and contrastive observations on scientific titles in written English and Spanish. *English for Specific Purposes*, n. 30, 124–137, 2011.

Vossen, P. (Ed.). *EuroWordNet General Document*. Version 3. University of Amsterdam, 2002. Disponível em: http://www.vossen.info/docs/2002/EWNGeneral.pdf. Acesso em: 12 de dez de 2019.

Vu, T. ; Aw, A. T.; Zhang, M. Term Extraction Through Unithood And Termhood Unification. *Proceedings of the Third International JCNLP Conference*, n. 2, p. 631–636, 2008.

Wang Y.; Melton G. B.; Pakhomov S. It's about this and that: a description of anaphoric expressions in clinical Text. *AMIA Annual Symposium Proceedings*, p. 1471–1480, 2011.

Wüster, E. *Introducción a la teoría general de la terminología y la lexicografía terminológica*. Barcelona: Universitat Pompeu Fabra, 1998.

Anexo

Tabela 2: Os 20 nomes, verbos e adjetivos mais frequentes dos *corpora* CORPORART-PT/IT

	Nomes		Verbos		Adjetivos	
	PT	IT	PT	IT	PT	IT
1	**arte**	**arte**	**ser**	**essere**	**Público**	**Pubblico**
2	**espaço**	**opera**	ir	**potere**	**Urbano**	**Urbano**
3	**obra**	**artista**	**ter**	**fare**	**Novo**	**Artístico**
4	**cidade**	**spazio**	**poder**	**avere**	**artístico**	Stesso
5	**forma**	**città**	**fazer**	**realizzare**	**social**	**Nuovo**
6	**artista**	**progetto**	estar	dovere	**grande**	**Culturale**
7	escultura	**luogo**	haver	**creare**	próprio	**Sociale**
8	monumento	**anno**	**dar**	**dare**	**primeiro**	**Diverso**
9	**projeto**	**parte**	dever	diventare	**cultural**	**Primo**
10	**trabalho**	art	**ver**	**trovare**	**diferente**	**Grande**
11	lisboa	**forma**	referir	portare	municipal	Street
12	elemento	street	**encontrar**	nascere	**político**	**contemporâneo**
13	relação	caso	**considerar**	volere	português	Tale
14	**intervenção**	**lavoro**	passar	vedere	maior	Vero
15	**ano**	**intervento**	apresentar	**considerare**	visual	**politico**
16	vez	modo	**realizar**	utilizzare	estético	possibile
17	**lugar**	tempo	partir	mettere	**contemporâneo**	ultimo
18	**parte**	art	dizer	andare	nacional	locale
19	local	interno	**criar**	prendere	importante	italiano
20	processo	diritto	vir	porre	diverso	creativo

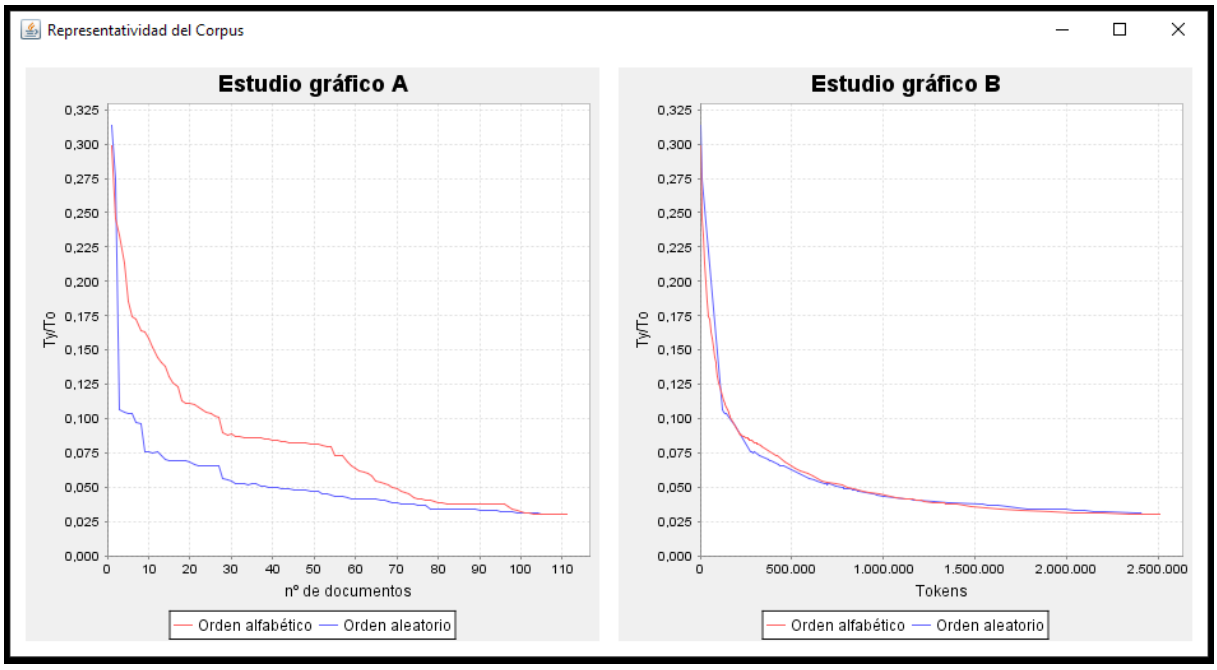

Figura A: Resultados para o CORPORART.PT, ReCor

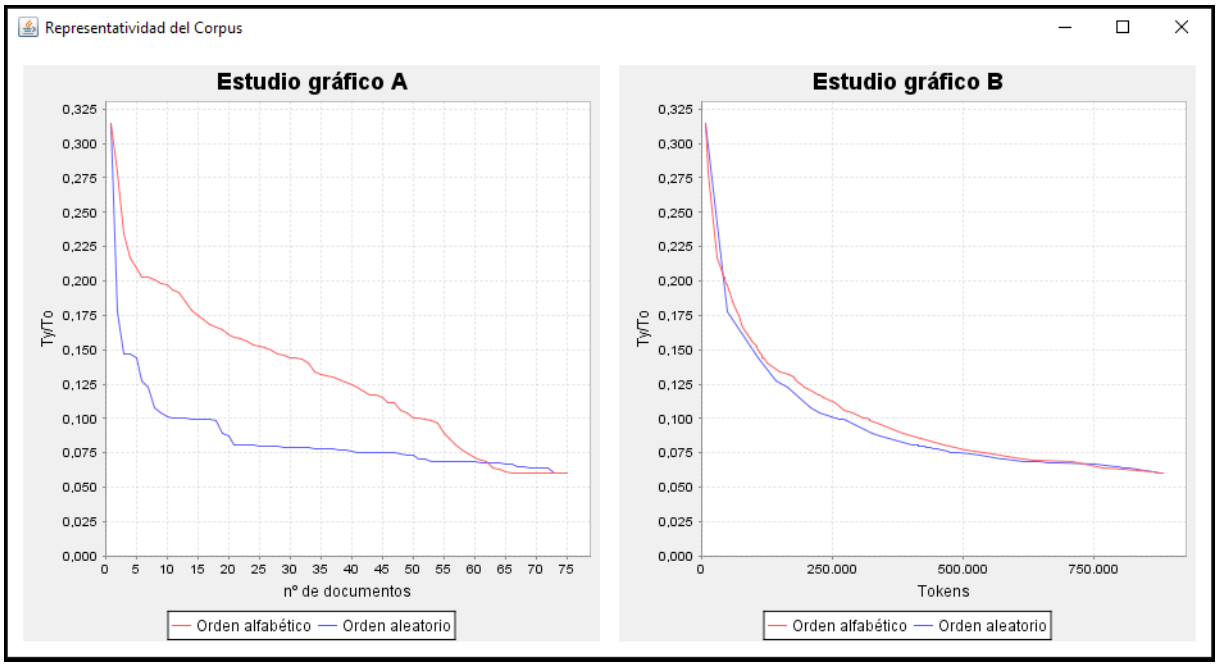

Figura B: Resultados para o CORPORART.IT, ReCor

Recebido: 18/12/2019.
Aprovado: 18/02/2020.