**Universidade do Minho**
Escola de Letras, Artes e Ciências Humanas

Iván Arias Arias

# Anotação semântica (semi)automática de corpora: a frase nominal em alemão

junho de 2022

Anotação semântica (semi)automática de corpora:
a frase nominal em alemão

Iván Arias Arias

UMinho | 2022


Europäischer Master für Lexikographie
EMLex
European Master in Lexicography
With the support of the
Erasmus+ Programme
of the European Union

**Universidade do Minho**
Escola de Letras, Artes e Ciências Humanas

Iván Arias Arias

# Anotação semântica (semi)automática de corpora: a frase nominal em alemão

Dissertação de Mestrado
Mestrado Europeu em Lexicografia

Trabalho efetuado sob a orientação do(a)
**Professor Doutor Álvaro Iriarte Sanromán**
**Professora Doutora María José Domínguez Vázquez**

junho de 2022

*À minha bisavó Remédios,*

in memoriam

## AGRADECIMENTOS

Gostaria de exprimir a minha gratidão a todas as pessoas que, de algum modo, me acompanharam ao longo do meu percurso académico e que ajudaram a que esta dissertação se concretizasse.

Em primeiro lugar, quero agradecer ao Professor Doutor Álvaro Iriarte Sanromán, por ter aceitado ser o meu orientador, por ter acreditado nas minhas capacidades para concluir esta tese no tempo previsto e por dirigir a realização desta tese com o seu enorme rigor científico. Obrigado ainda por me ter ensinado tanto sobre *mel'čukadas*, pois foi isso que motivou, em grande medida, a minha propensão pela semântica lexical.

Agradeço, de igual forma, à Professora Doutora María José Domínguez Vázquez, por me ter assistido durante a elaboração desta dissertação enquanto orientadora, por me ter acompanhado ao longo do meu percurso académico universitário e por me dar sempre tão bons conselhos. Ficarei grato *ad aeternum* por todas as horas de conversas e discussões que sempre acabam por levar a resultados *prometedores*.

À Professora Doutora Idalete Maria da Silva Dias, diretora do Mestrado em Lexicografia na Universidade do Minho, obrigado por fazer com que em Braga me sentisse como em casa, por todo o seu apoio nos momentos mais difíceis e por me receber sempre com um sorriso.

Ao Mestrado Europeu em Lexicografia (EMLex), por me dar esta incrível oportunidade que contribuiu não só para a minha formação académica, mas também para o meu crescimento pessoal.

À Margarida Castro, por ser uma amiga incondicional, por saber como lidar comigo em momentos de stress, por ter a paciência de Jó... Esta experiência nunca teria sido a mesma sem o seu respaldo. Mais queria desejar à Margarida boa sorte e muito sucesso! *O que EMLex uniu ninguém o separa!*

Às minhas amigas, agradeço por todos os magníficos momentos que passámos juntas, por todo o incentivo para perseguir os meus sonhos e por me apoiarem sempre.

À minha família, agradeço por me terem encorajado a estudar algo que me fizesse feliz.

Ao meu irmão, por ser o meu maior orgulho.

A todas, os meus mais sinceros agradecimentos...

## DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

## RESUMO

### Anotação semântica (semi)automática de *corpora*: a frase nominal em alemão

Nos dias de hoje, no âmbito da investigação e da prática lexicográfica, a utilização de *corpora* tem-se revelado muito recorrente, principalmente pelo facto de ser considerada como a metodologia mais fiável para alcançarmos exemplos representativos das línguas naturais. Embora as ferramentas de Processamento de Língua Natural (PLN) tenham conseguido grandes avanços na anotação morfossintática de textos, continua a faltar uma anotação semântica exaustiva e sistematizada. Esta carência evidencia-se principalmente quando se fala em lexicografia e gramática de valências, pois na bibliografia teórica (cf. Domínguez, 2011) aponta-se para o facto de a valência semântica ser fulcral para a delimitação de argumentos que acompanham um lexema considerado como portador de valência. Daí surge, no contexto desta investigação, a necessidade de uma aproximação à anotação semântica de *corpora*, em que se preste atenção especial aos argumentos no nível da frase nominal e ao seu comportamento semântico, para além da etiquetagem morfossintática com a qual contamos normalmente. A gramática e lexicografia de valências, assim como a semântica léxica, constituem, portanto, o ponto de partida teórico da presente dissertação de mestrado. No que diz respeito à metodologia, o presente trabalho cingir-se-á à análise das estruturas argumentais de três nomes do campo semântico da comunicação em alemão (*Bericht*, *Diskussion* e *Frage*) e, através de metodologia de PLN, desenhar-se-á um *API script* que possibilite o cruzamento de dados de *corpora* com alguns pacotes lexicais delimitados e criados no âmbito dos projetos PORTLEX, MultiGenera e MultiComb. Esta metodologia permitir-nos-á analisar, *a posteriori*, a fiabilidade do *script* desenvolvido, e conduzirá para a extração de conclusões relativas ao valor que poderia trazer consigo a anotação semântica sistematizada de *corpora*.



**Palavras-chave:** anotação semântica, *corpora*, pacote lexical, PLN, valência nominal

## ZUSAMMENFASSUNG

### (Semi)automatische semantische Annotation von Korpora: die Nominalphrase im Deutschen

Heutzutage wird in der Wörterbuchforschung und in der Lexikographie immer häufiger auf Korpora zurückgegriffen, weil sie als zuverlässige Methode gelten, um repräsentative Beispiele der natürlichen Sprache zu finden. Obgleich die Entwicklung von Tools im Bereich der natürlichen Sprachverarbeitung (NLP) dazu führte, dass die Texte morphosyntaktisch annotiert sind, fehlt es immer noch an einer umfassenden und systematisierten semantischen Annotation. Dieser Mangel wird besonders deutlich, wenn man sich mit der Valenzlexikographie und der Valenzgrammatik befasst, da in der Literatur (vgl. Domínguez, 2011) darauf hingewiesen wird, dass die semantische Valenz wesentlich für die Abgrenzung von Ergänzungen ist, die neben einem als Valenzträger zu betrachtenden Lexem auftreten. Daraus ergibt sich, dass es einem Ansatz zur semantischen Annotation von Korpora bedarf, bei dem die nominalen Ergänzungen und ihr semantisches Verhalten im Vordergrund stehen und der sich zum Ziel setzt, die Grenzen der bereits existierenden morphosyntaktischen Annotation zu überschreiten. Die Valenzgrammatik und -lexikographie sowie die lexikalische Semantik stellen daher den theoretischen Ausgangspunkt der vorliegenden Masterarbeit dar. Die Vorgehensweise dieser Arbeit beschränkt sich auf die Analyse der Argumentstrukturen von drei Substantiven aus dem semantischen Feld der Kommunikation im Deutschen (*Bericht*, *Diskussion* und *Frage*). Mithilfe von Tools der NLP wird ein Skript entwickelt, das einen Abgleich zwischen den aus Korpora stammenden Daten und den lexikalischen Paketen entnommenen Daten ermöglicht. Die sog. lexikalischen Paketen wurden im Rahmen der Projekte PORTLEX, MultiComb und MultiGenera erstellt. Anschließend ist die Zuverlässigkeit des erstellten Skripts zu analysieren und es werden Schlussfolgerungen hinsichtlich des Wertes der systematisierten semantischen Annotation von Korpora gezogen.



**Schlüsselwörter:** semantische Annotation, Korpora, lexikalisches Paket, NLP, nominale Valenz

# ÍNDICE

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS ..... ii

AGRADECIMENTOS ..... iii

DECLARAÇÃO DE INTEGRIDADE ..... iv

RESUMO ..... v

ZUSAMMENFASSUNG ..... vi

LISTA DE FIGURAS ..... ix

LISTA DE TABELAS ..... x

Introdução ..... 1

Objetivos da investigação ..... 4

Enquadramento teórico ..... 7

3.1. O campo lexical da comunicação ..... 7

3.2. A valência nominal ..... 10

Investigações relacionadas ..... 18

4.1. FrameNet ..... 18

4.2. Corpus Pattern Analysis (CPA) ..... 19

4.3. Os projetos PORTLEX, MultiGenera e MultiComb ..... 20

Enquadramento metodológico ..... 31

5.1. Perguntas de investigação ..... 31

5.2. Metodologia ..... 33

Apresentação e análise dos resultados ..... 42

6.1. O substantivo Bericht ..... 42

6.2. O substantivo Diskussion ..... 56

6.3. O substantivo Frage ..... 68

Conclusão ..... 81

Referências bibliográficas ..... 85

a. Literatura científica ..... 85

b. Recursos em linha ..... 89

Anexos ..... 91

## LISTA DE FIGURAS

Figura 1: Objetivos gerais e objetivos específicos ..... 6
Figura 2: Captura de ecrã do artigo lexicográfico *Diskussion* em PORTLEX ..... 22
Figura 3: Captura de ecrã de *Xera* ..... 27
Figura 4: Captura de ecrã de *Xera* ..... 28
Figura 5: Captura de ecrã de *Combinatoria* ..... 29
Figura 6: Captura de ecrã de *CombiContext* ..... 30
Figura 7: Fases de trabalho metodológico ..... 33
Figura 8: Frequência absoluta em Sketch Engine: German Web 2018 das estruturas selecionadas ..... 49
Figura 9: Frequência absoluta das ocorrências anotadas como {animado, humano} para *Bericht* ..... 51
Figura 10: Número aproximado de ocorrências {animado, humano} no *corpus* ..... 53
Figura 11: Frequência absoluta em Sketch Engine: German Web 2018 das estruturas selecionadas ..... 62
Figura 12: Frequência absoluta das ocorrências anotadas como {animado, humano} para *Diskussion* ..... 64
Figura 13: Número aproximado de ocorrências {animado, humano} no *corpus* ..... 66
Figura 14: Frequência absoluta em Sketch Engine: German Web 2018 das estruturas selecionadas ..... 74
Figura 15: Frequência absoluta das ocorrências anotadas como {animado, humano} para *Frage* ..... 76
Figura 16: Número aproximado de ocorrências {animado, humano} no *corpus* ..... 78

## LISTA DE TABELAS

Tabela 1: Visão geral do conflito terminológico para actantes e circunstantes ..... 12
Tabela 2: Lista de nomes extraídos do pacote lexical {animado, humano} ..... 38
Tabela 3: Estrutura argumental do nome *Bericht* ..... 44
Tabela 4: Dados mais frequentes para a estrutura *Bericht* + genitivo em Sketch Engine: German Web 2018 ..... 50
Tabela 5: Estrutura argumental do nome *Diskussion* ..... 58
Tabela 6: Dados mais frequentes para a estrutura *Diskussion* + genitivo em Sketch Engine: German Web 2018 ..... 63
Tabela 7: Estrutura argumental do nome *Frage* ..... 70
Tabela 8: Dados mais frequentes para a estrutura *Frage* + genitivo em Sketch Engine: German Web 2018 ..... 75

# 1

## Introdução

Nas últimas décadas, alguns estudos, designadamente a publicação de Schierholz (2008), têm apontado para a necessidade de operacionalizar a descrição de estruturas argumentais ou padrões valenciais. Para tal fim, o recurso a *corpora* linguísticos desempenha um papel essencial, já que se trata de um método de recolha de dados linguísticos muito fidedigno (Teubert, 2001) devido ao facto de representar usos de língua natural e não criados *ad hoc*. Aliás, deve ser salientada, em consequência, a possibilidade de integrar posteriormente estes exemplos reais, por exemplo, para a delimitação de estruturas argumentais em bases de dados lexicográficas e dicionários.

De forma simultânea, alguns autores (*vd.* Iriarte, 2001, p. 18) chamam a atenção para a dificuldade de realizar análises linguísticas e de combinatória lexical partindo somente de traços semânticos inerentes a unidades lexicais. Não obstante, deve-se ter em conta que, por vezes, ocorrem estruturas idênticas do ponto de vista formal ou morfossintático e de combinação livre, cuja diferença única reside no plano semântico (vejam-se as duas possibilidades listadas para o mesmo predicado nominal a seguir: *a discussão* {*dos estudantes* | *das tarefas*}). Nesses casos, a desambiguação em *corpora* só é possível através da aplicação de traços semânticos.

Para a análise destes fenómenos linguísticos, Zgusta (1971) centra o foco na importância de trabalhar com enunciados concretos que nos permitam deduzir o significado de determinadas unidades lexicais, pois o sentido lexical nem sempre pode ser inferido de palavras isoladas como formando parte do sistema abstrato da língua. Neste sentido, para a realização da presente pesquisa, parte-se da hipótese de que a frase nominal é a unidade mínima com significado, pela necessidade de contar com um contexto para a delimitação de um sentido (*vd.* Gross, 2013). Com base neste pressuposto, estabelece-se o contexto frasal como âmbito de aplicação para esta dissertação, sendo que esta delimitação será já essencial para o enquadramento teórico do trabalho.

Aliás, de modo a atingir uma melhor descrição das relações lexicais intralinguísticas, a anotação exaustiva de *corpora* com a inclusão de parâmetros semânticos, sejam eles categoriais ou relacionais (Engel, 2004), converte-se em *conditio sine qua non*. Pretende-se, deste modo, preencher uma lacuna na etiquetagem de *corpora* linguísticos, pois principalmente para a análise de padrões valenciais, a

anotação semântica é fundamental (*vd.* Domínguez, 2014a; Domínguez *et al.*, 2018; López, 2020) e ainda não está incluída com informação categorial e/ou relacional em *corpora*.

A seguir, apresenta-se a estruturação deste trabalho.

No capítulo 2 pretende-se estabelecer quais são os objetivos gerais e os objetivos específicos que articularão a realização da presente pesquisa, sendo que se parte da tese de que os *corpora* com apenas anotação morfossintática podem originar alguns problemas na altura da recolha de dados linguísticos. Assim, também se defenderá a importância de integrar técnicas de Processamento de Linguagem Natural (PLN) no âmbito da linguística aplicada e da lexicografia.

A epígrafe 3 cinge-se a aspetos teóricos, de modo que a teoria dos campos lexicais (Geckeler, 1971; Coseriu, 1977) desempenhará um papel fundamental inicialmente, pois cabe delimitar o campo lexical do que fazem parte as unidades lexicais que serão analisadas. Por outra parte, pretende-se alicerçar a presente pesquisa sobre a gramática de valências, nomeadamente na valência nominal, pelo facto de considerarmos a frase nominal como sendo a unidade mínima de significado. Destarte, oferecer-se-á uma visão geral sobre os fundamentos conceituais da valência nominal enquanto sistema *sui generis* (Teubert, 1979).

Outras abordagens para o tratamento da semântica combinatória serão introduzidas no capítulo 4, onde se fornecerá informação sobre diferentes possibilidades para o tratamento do léxico. Em primeiro lugar, falar-se-á da teoria de cenários (*frames*) de Fillmore (1982, 1985). Depois, centrar-se-á o foco na abordagem de Hanks (2004, 2013), conhecida como *Corpus Pattern Analysis* (CPA). Por último, tenciona-se abordar mais pormenorizadamente o dicionário PORTLEX e outras ferramentas multilingues desenhadas para a manipulação de dados linguísticos no âmbito dos projetos MultiComb e MultiGenera. Quer o dicionário PORTLEX quer as ferramentas referidas constituem um antecedente indiscutível para a conceção desta pesquisa.

Em seguida, no capítulo 5, visa-se delimitar a pergunta de investigação sobre a qual se baseará a análise posteriormente realizada. Nesta linha, a demarcação de uma questão de investigação com as suas correspondentes hipóteses considerar-se-á essencial para a execução de um trabalho linguístico ou lexicográfico empírico. Igualmente, apresentar-se-á a metodologia desenhada propositadamente (a criação de um *script* informático *ad hoc*) para atingir os objetivos perseguidos na presente dissertação.

Na secção 6 aparecem em primeiro plano os resultados provenientes da análise realizada, o que nos permitirá avaliar a viabilidade da metodologia desenhada e a sua futura aplicação para outros projetos. Realizar-se-á, para tal propósito, uma análise de tipo quantitativo baseada em critérios de

frequências, combinada com a retirada de conclusões qualitativas. Deste modo, pretende-se encerrar a discussão aberta com as perguntas de investigação que se formularão.

Concluir-se-á o trabalho com um resumo dos resultados obtidos na pesquisa, assim como com uma síntese sobre a viabilidade ou não-viabilidade de aplicar o *script* informático desenhado a maiores quantidades de texto em língua alemã. Refletir-se-á, igualmente, sobre a necessidade de introduzir anotação semântica em *corpora* linguísticos, o que evidencia, aliás, em termos gerais, a importância da semântica e da lexicologia, que costumam ser relegadas para segundo plano na linguística computacional ou de *corpus*, pois as consultas morfossintáticas em *corpora* são implementadas de forma massiva em detrimento, por vezes, da etiquetagem semântica.

# 2

## Objetivos da investigação

> „Eine wichtige Aufgabe der Lexikografie der digitalen Zukunft ist die geordnete Zusammenführung von automatisch aus Textkorpora erzeugten und gezielt aufbereiteten Daten sowie einer benutzerorientierten Präsentation. Die gesellschaftliche Relevanz solcher Informationssysteme wird gefestigt, wenn die zugrundeliegenden Korpora das gesamte sprachliche Diasystem spiegeln und für ForscherInnen frei zur Verfügung stehen." (Mestrado Europeu em Lexicografia, 2018[1])

Tal como se pode ler na citação anterior, que corresponde à tese *Villa Vigoni* número 7, um dos desafios principais para a lexicografia da atual era digital consiste na anotação exaustiva e na representação precisa de diassistemas linguísticos em *corpora*. Para atingir este objetivo, torna-se fulcral, do meu ponto de vista, incluir informação semântica nos diferentes recursos de análise e consulta. Assim, para além dos dados que se podem extrair de *corpora*, tais como os formais (as consultas CQL, por exemplo) ou os quantitativos (entre os que se destacam os critérios de frequência), uma etiquetagem semântica pode contribuir para uma análise mais fidedigna das unidades léxicas que estão recolhidas e processadas nos *corpora* com os quais se trabalha. Corrobora-se, aliás, que a falta desta anotação semântica pode conduzir, dependendo do tipo de pesquisa pretendida, para uma maior dificuldade quando se pretendem desambiguar resultados obtidos através de uma consulta em *corpora*, pois, por vezes, a anotação morfossintática torna-se insatisfatória. Por exemplo, os critérios de frequência não são sempre suficientes para o desenho de alguns recursos, entre os quais cabe ressaltar aqueles que visam descrever estruturas argumentais (cf. Domínguez *et al.*, 2018; López, 2020).

Na área da lexicografia e da lexicologia, diferentes autores, entre os que se destaca Wierzbicka (1996), apontam para as dificuldades de desambiguação em contextos de polissemia e acabam por apresentar a polissemia como um problema potencial para a extração de dados linguísticos em várias ocasiões. Para paliar alguns dos efeitos negativos derivados da polissemia, visa-se entender, no contexto deste projeto, a frase nominal como sendo a unidade mínima de significado, pelo facto de só assim ser

[1] As teses *Villa Vigoni* foram redigidas em 2018 após uma reunião de especialistas na área da lexicografia cujo objetivo era analisar os desafios da investigação lexicográfica no futuro próximo. Para mais informações, consulte-se a bibliografia *infra*.

conferido um sentido co[n]textual a nível frasal, para além do sentido inerente e intrínseco dos lexemas que serão tratados (cf. Gross, 2013, p. 31).

A consulta de *corpora* através de CQL[2] morfossintáticas não acaba por atenuar esse problema, pois a partir da utilização, por exemplo, da CQL `[lemma="Diskussion"] [tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"][tag="N.*"]` obtêm-se resultados como *Diskussion der Kandidaten* (> *discussão dos candidatos*), em que a frase em genitivo corresponde à realização superficial do complemento sujeito (*os candidatos discutem*), ou como *Diskussion der Ergebnisse* (> *discussão dos resultados*), onde a frase em genitivo diz respeito ao papel semântico daquilo que é afetado pela ação ('discute-se sobre os resultados'). Na presente dissertação, centraremos o foco na realização superficial, morfossintática e semântica do primeiro actante ou, sendo mais específicos, do complemento sujeito (cf. Domínguez, 2014b), por ser ele o que transmite a informação relativa ao agente – aquele que realiza uma dada ação descrita no predicado –, embora sejamos conscientes de que se possa tratar, na maior parte dos casos, de estruturas poliargumentais, isto é, estruturas que contam com mais de um argumento no seu esquema actancial.

A partir do exposto, deduz-se que o objetivo principal da presente dissertação consiste em fazer uma aproximação à anotação semântica de *corpora* de modo a facilitar a consulta de grandes quantidades de texto evitando, com este mecanismo, algumas situações altamente ambíguas, como as frases preposicionais com *de* em português ou as frases em genitivo em alemão. Destarte, pretende-se desenvolver e avaliar uma ferramenta (fundamentalmente, um *script* informático) que permita uma anotação de *corpora* semiautomática, já que uma etiquetagem exaustiva realizada de forma manual requer mais tempo e um maior esforço por parte da equipa responsável pela execução das tarefas. Aliás, tenciona-se dar um contributo para o estabelecimento de aplicações de Processamento de Língua Natural (PLN), sendo que se recorre a métodos de análise linguística próprios desta perspetiva (cf. Gross, 2013). Sublinha-se que não só se pretende contribuir com um *script* original que possibilite uma anotação semântica automática de *corpora*, ao nível da frase nominal, senão que também se recorrerá a outros recursos já disponibilizados e desenhados no âmbito de PLN.

Na literatura alemã, autores como Wiegand (1998) ou Teubert (2001), salientam, de mais a mais, a importância do uso de *corpora* para atingir uma maior e melhor representatividade da língua natural e real. Embora eles não falem na necessidade de anotação semântica de forma explícita,

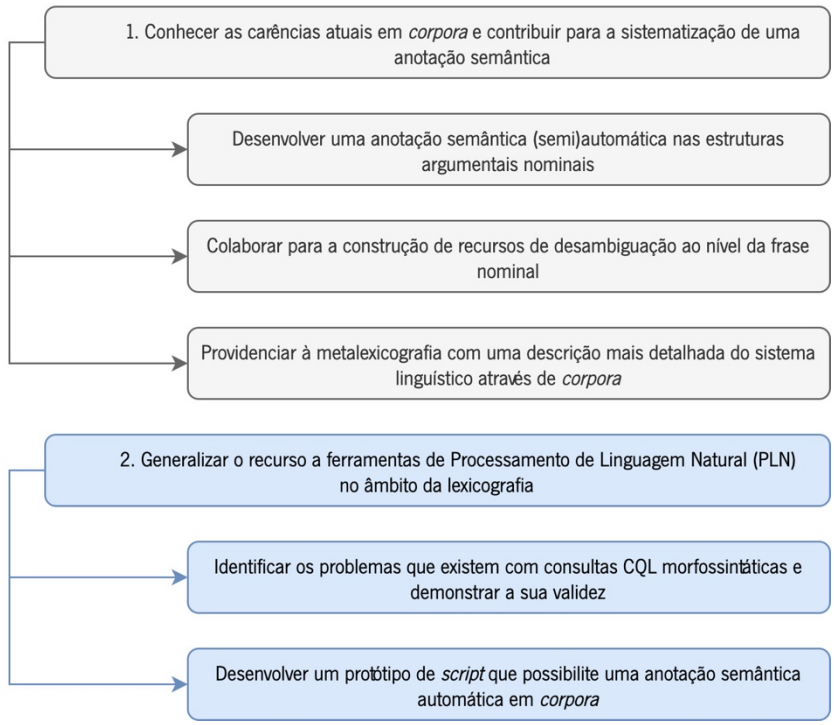
[2] Jakubícek *et al.* (2010, p. 742) afirmam que "a CQL query is a pattern which may match a token or series of tokens in the corpus. Each token is assigned a set of attributes (word form, lemma, part-of-speech tag etc.) and each corpus might be assigned a set of structures."

Domínguez (2014a) já aponta para os problemas que pode acarretar eventualmente a escasseza notória de etiquetagem semântica em *corpora*. Aliás, esta anotação diz respeito a dois níveis: primeiramente, ao nível frasal e contextual, fortemente ligado às estruturas argumentais e à gramática valencial, tratando-se do significado relacional (Engel, 2004; Domínguez, 2014a); por outro lado, ao nível do significado categorial e ontológico das unidades lexicais, que é este o que se revela como ponto de partida para análises semânticas ulteriores. Outros estudos executados de uma perspetiva similar, como o de López (2020), destacam igualmente a falta de uma anotação semântica sistematizada em *corpora* para a extração de informação válida.

De modo a conseguir os objetivos *supra* elencados, será preciso fornecer um enquadramento teórico bem delimitado para a presente dissertação, sendo que a valência nominal se tornará o nosso pilar fundamental. Depois de uma introdução teórica, cabe apresentar uma metodologia que permita atingir estes alvos, cujos resultados serão apresentados na última secção (§ 6) deste trabalho. Em suma, a figura 1 ilustra, de forma sucinta, os principais objetivos gerais e específicos da presente tese de mestrado.

**Figura 1**

*Objetivos gerais e objetivos específicos*




*Nota*. Nesta figura de elaboração própria listam-se os dois objetivos principais que se pretendem alcançar no âmbito da presente dissertação, assim como alguns dos objetivos específicos subjacentes às motivações para a realização do presente trabalho.

# 3

## Enquadramento teórico

Neste capítulo serão abordados os fundamentos teóricos que servirão como ponto de partida para a análise que será efetuada na presente dissertação. Para tal finalidade, apresentar-se-ão algumas noções fundamentais relativamente à teoria dos campos lexicais, de modo a conseguirmos uma delimitação clara do âmbito de trabalho, dado que os substantivos que se analisarão na presente investigação são, entre si, co-hipónimos do campo lexical da comunicação (*vd.* subcapítulo 3.1.). Em seguida, o foco do trabalho centrar-se-á na gramática de valências, sendo que nos debruçaremos fundamentalmente sobre a valência nominal, pelo facto de serem três nomes o ponto de partida básico da presente pesquisa (*vd.* subcapítulo 3.2.). A gramática de valências acabará, então, por se revelar como parte essencial e como alicerce para o enquadramento metodológico e para a análise posterior.

### 3.1. O campo lexical da comunicação

Antes de estabelecermos uma delimitação exaustiva do campo lexical da comunicação, parece necessário chamarmos a atenção para aquilo que entendemos com "comunicação". Conforme a segunda aceção do *Dicionário da Língua Portuguesa* (2009) da Porto Editora, a unidade lexical "comunicação" é definida como "troca de informação entre indivíduos através da fala, da escrita, de um código comum ou do próprio comportamento". Evidencia-se, através da paráfrase de significado *supra* citada, que o modelo de comunicação proposto por R. Jakobson (1971) na linguística teórica tradicional desempenha ainda um papel fundamental para a delimitação do campo semântico descrito no presente trabalho. Destarte, é importante para a seleção de substantivos termos em consideração que no significado deles deve estar presente a seguinte informação lexical-cognitiva: deve tratar-se sempre da transmissão de uma mensagem (Z) de um emissor (pessoa A) para um recetor (pessoa B). Assim sendo, trabalhar-se-á fundamentalmente com substantivos cujo significado implique a existência e/ou presença de, quando menos, um interlocutor e uma mensagem. Para além disso, tenciona-se prosseguir com a pesquisa iniciada em Arias (2020) sobre substantivos do campo lexical da comunicação em língua alemã.

No que diz respeito à teoria semântica de campos lexicais, deve-se considerar, antes de mais, que existe frequentemente na literatura científica e teórica uma problemática em torno da fixação de uma nomenclatura homogénea para designar esta realidade. Embora estas contrariedades terminológicas não sejam indispensáveis para o presente trabalho, interessa apontarmos para a existência de diferentes

conceitos que designam a mesma realidade linguística. Deste modo, na bibliografia científica, principalmente em língua alemã, encontram-se várias designações: Lutzeier (1981, p. 85) ou Lyons (1977, p. 207) ressaltam conceitos como *Wortfeld*, *Bedeutungsfeld*, *Sinnbezirk*, *lexikalisches Feld*, *sprachliches Feld* ou *Begriffsfeld* para fazerem alusão aos campos léxicos ou semânticos.

Aliás, convém esclarecer que, na investigação linguística em língua portuguesa, costuma-se empregar o termo "campo lexical" para fazer referência a um "conjunto de palavras associadas, pelo seu significado, a um determinado domínio conceptual" (Dicionário Terminológico[3]), enquanto a noção de "campo semântico" diz respeito ao "conjunto de significados que uma palavra pode ter nos diferentes contextos em que se encontra". No presente trabalho, e de acordo com Coseriu (1977, pp. 170-171), utilizar-se-á a designação de "campo lexical" e partir-se-á da seguinte definição:

> "Un *campo léxico* es una estructura paradigmática constituida por unidades léxicas que se reparten una zona de significación común y que se encuentran en oposición inmediata las unas con las otras. [...] Se trata siempre de unidades léxicas entra las que existe «selección» (posibilidad de elección) en un punto determinado de la cadena hablada. [...] Puede decirse también que un campo léxico está constituido por el término presente en un punto determinado de la cadena hablada y los términos que su presencia excluye de manera inmediata."

A possibilidade de escolha entre as várias unidades lexicais que não se dão simultaneamente na cadeia falada, ou seja, que se encontram dentro do plano paradigmático ou associativo e não do sintagmático (cf. Saussure, 1986, pp. 207-208), permite-nos afirmar que se trata de unidades que são consideradas como co-hipónimos. Em consequência, para alguns dos lexemas com os quais se trabalhará pode-se propor o substantivo "comunicação" como hiperónimo comum, embora isto não impeça que outros nomes possam funcionar outrossim como hiperónimos[4]. Daí deduz-se, consoante o defendido por Saussure (1986, p. 196), que, no eixo paradigmático, "todas as palavras que exprimem ideias vizinhas [se] limitam reciprocamente". Relativamente ao campo lexical da comunicação, verifica-se que unidades lexicais como *debate*, *discussão*, *conversação* ou *relatório* estão delimitadas no plano da significação graças à presença de sememas específicos no significado de algumas delas.

Não obstante, a delimitação de campos lexicais não se revela frequentemente como tarefa fácil, pois devem-se considerar duas questões fundamentais:

---

[3] Estes termos podem ser consultados no Dicionário Terminológico (*vd.* bibliografia), disponibilizado em linha e orientado para professores de português do ensino básico e secundário com termos de carácter linguístico.

[4] Para uma esquematização de algumas relações paradigmáticas no campo semântico da comunicação, consulte-se Arias (2020, p. 10).

> "Die Frage der Abgrenzungen ist eine doppelte: sie betrifft einmal die Binnengrenzen des Feldes, d.h. die inhaltlichen Abgrenzungen der einzelnen Feldglieder untereinander, zum anderen betrifft sie die Außengrenzen des Feldes, d.h. seine Abgrenzung gegenüber benachbarten Feldern." (Geckeler, 1971, p. 193)

Por um lado, devemos analisar os limites externos (*Außengrenzen*) do campo lexical, que podem ser facilmente delimitados através da paráfrase de significado que referimos anteriormente. Neste sentido, serão considerados, portanto, como candidatos potenciais do campo lexical da comunicação aqueles substantivos cujo sentido tenha a ver com a transmissão de uma mensagem (Z) de um emissor (A) para um recetor (B). Por outro lado, pode ser mais difícil demarcar os limites internos (*Binnengrenzen*), pois por vezes a determinação tem de ser realizada com o recurso a sememas muito específicos.

De modo a delimitarmos internamente o campo lexical da comunicação, consideraremos, de acordo com Arias (2020) e com Schumacher (1986), o semema [+monológico] como sendo oposto a [+dialógico] e como sendo aquele que permite uma separação entre os lexemas pertencentes ao campo lexical da comunicação no nível superior. No que diz respeito ao primeiro grupo, o traço semântico [+monológico], tal como se pode inferir da sua etimologia, implica que a produção linguística acontece só num sentido: de emissor para recetor. Fazem parte deste grupo (Arias, 2020, p. 7) palavras como *Behauptung* (> *afirmação*), *Bericht* (> *relatório*) ou *Erzählung* (> *narração*).

Por sua parte, para a consolidação do semema [+dialógico], podemos citar Schumacher (1986, p. 667), segundo quem "zwei oder mehr Gesprächsbeteiligte in einer Gesprächssituation die Rollen des Sprechers und Hörers wechseln." Em consequência, trata-se sempre de lexemas cujo sentido tem a ver com o facto de uma troca comunicativa, em que os participantes podem alterar, entre si, os papeis de emissor e recetor. Fazem parte deste grupo os seguintes substantivos: *Debatte* (> *debate*), *Diskussion* (> *discussão*) e *Gespräch* (> *conversa*),

Com esta descrição da teoria dos campos lexicais, e com o foco centrado no campo lexical da comunicação, pretende-se fornecer uma visão geral sobre o conteúdo proposicional dos nomes que serão apurados para a presente pesquisa, sendo que se estabelece uma distribuição lexical destes nomes perante a existência de outros substantivos do mesmo campo lexical. Desta forma, consegue-se delimitar o contexto de trabalho em que será aplicada depois a metodologia selecionada. Aliás, o facto de realizarmos a pesquisa para unidades lexicais próximas no que concerne ao seu significado, permitir-nos-á retirar conclusões mais objetivas acerca da validade e utilidade de anotação semântica de *corpora*.

## 3.2. A valência nominal

Como já foi referido, a gramática de valências é um dos pilares fundamentais da presente investigação. Antes de nos debruçarmos sobre os princípios linguísticos subjacentes a esta teoria e que serão fulcrais para a análise metodológica ulterior, parece relevante destacar que o seu precursor foi o francês L. Tesnière (1959[5]), quem defendia a seguinte hipótese, entre outras que serão mencionadas mais pormenorizadamente ao longo do presente capítulo:

> "The verb may therefore be compared to a sort of **atom**, susceptible to attracting a greater or lesser number of actants, according to the number of bonds the verb has available to keep them as dependents. The number of bonds a verb has constitutes what we call verb's **valency**." (Tesnière, 2015, p. 239)

Deste modo, observa-se que a gramática de valências foi inicialmente aplicada somente para analisar o comportamento sintático-semântico de verbos, sendo que o verbo funciona como um "átomo" que provoca a ocorrência obrigatória de outros elementos à sua beira. O verbo, tal como os átomos na química, exerce uma força de atração sobre outros elementos que aparecem no contexto oracional. Baseando-se na teoria proposta por Tesnière, no âmbito da linguística estrutural, autores como Zifonun (2003, p. 355) definem a valência como „die Fähigkeit eines Valenzträgers[6], in beliebigen Sätzen seiner Umgebung eben jene strukturellen Beziehungen aufzuprägen. [...] Valenz [ist] in jedem Fall als Form von Abhängigkeit zu interpretieren." Do ponto de vista mais tradicional, costuma distinguir-se, portanto, entre valência sintática, semântica e lógico-pragmática, embora esta conceção tenha sido alargada nas últimas décadas:

> „Seit den Anfängen der Valenztheorie hat sich der Valenzbegriff zu einem Multimodulkonzept entwickelt, wo unterschiedliche als komplementär anzusehende nicht isomorphe Ebenen Eingang finden. Klassische Einteilungen der Valenz weisen auf eine syntaktische, eine semantische und eine logisch-pragmatische Valenz hin." (Domínguez, 2011, p. 17)

A conceção multidimensional da valência aponta para mais níveis de descrição, assim como para a consideração de outras classes de palavras, além do verbo, como portadores de valência. Primeiramente, no que diz respeito aos níveis de descrição valencial (*Valenzebenen*), Helbig (1992, p. 7) fala na diferenciação entre a valência lógica, a semântica e a sintática. A valência lógica tem a ver com

[5] Embora o livro *Eléments de syntaxe structurale*, onde L. Tesnière explica os pormenores da teoria de valência, fosse publicado em 1959, no presente trabalho será citada principalmente a tradução para o inglês de 2015.
[6] Zifonun (2003) apresenta a valência como um conceito mais abrangente, pois a autora fala em "Valenzträger" (portadores de valência) e não se ocupa só da descrição verbal, tal como fazia tradicionalmente a gramática de valências seguindo Tesnière (1959).

o facto de as representações objetivas da realidade extralinguística serem transmitidas através de signos linguísticos, sendo que a estrutura fica condicionada pela interligação entre os aspetos extra- e intralinguísticos (cf. Helbig, 1992, p. 7). A valência semântica, por sua parte, corresponde à determinação dos papeis semânticos (agente, paciente...) e à delimitação de traços semânticos que devem ocorrer num determinado contexto linguístico, de forma que a característica [+humano] ou [-animado], por exemplo, não são intercambiáveis entre si. Helbig (1992, pp. 7-8) exemplifica isto com o caso do verbo alemão *bewundern* (> *admirar*):

> „Ein Prädikat wie *bewundern* hat zwei Leerstellen, die durch Argumente zu besetzen sind. Aber das als Subjekt fungierende Argument muß das Merkmal [Hum], kann nicht das Merkmal [-Anim] oder [Abstr] haben [...]. Umgekehrt kann das als Objekt fungierende Argument mehreren semantischen Klassen angehören, ist nahezu frei von Restriktionen in dieser Hinsicht."

Quanto à valência sintática, Helbig (1992, p. 9) afirma que ela corresponde fundamentalmente à obrigatoriedade ou facultatividade dos elementos que aparecem na proximidade de um determinado portador de valência, sendo que a quantidade de complementos e suplementos e a sua realização linguística desempenham um papel crucial neste nível. Domínguez (2011, p. 20) acrescenta o nível pragmático e defende que algum parâmetro valencial pode estar condicionado pela situação comunicativa em que ocorre, assim como pelos atos de fala.

Correlacionado com o nível sintático, e de forma a explicar a relação de dependência entre os constituintes oracionais e o verbo, Tesnière diferencia entre actantes e circunstantes. Por um lado, os actantes são "the beings or things, of whatever sort these might be, that participate in the process, even as simple extras or in the most passive way" (Tesnière, 2015, p. 97). Por outro lado, os circunstantes "express the circumstances of time, place, manner, etc. in which the process unravels" (Tesnière, 2015, p. 97). Contudo, atualmente podem-se encontrar, na literatura teórica, divergências no que diz respeito às designações utilizadas para falar da diferença entre actantes e circunstantes, uma vez que de acordo com Hölzner (2007, p. 1) "das Valenzkonzept hat sich seit Tesnières Zeit immer stärker etabliert und ist fortlaufend modifiziert und erweitert worden". Nesta linha, Storrer (2003, p. 766) remete para um esquema similar ao da tabela 1 para apresentar, de forma sucinta, uma parte da nomenclatura desta área.

**Tabela 1:**

*Visão geral do conflito terminológico para actantes e circunstantes*

| actants | circonstants |
|---|---|
| Ergänzungen | Angaben |
| Aktanten | Zirkumstanden |
| Mitspieler | Umstandsbestimmungen |
| Valenzpartner | (freie) Angaben |
| Komplemente | Supplemente |
| Argumente | Adjunkte |
| (pt.) actantes | (pt.) circunstantes |
| (pt.) complementos | (pt.) suplementos |
| (pt.) argumentos | (pt.) adjuntos |

*Nota.* Esta tabela foi elaborada pelo autor desta dissertação a partir de Storrer (2003, p. 766). Deve-se notar que foi modificada para a inclusão dos termos em português inexistentes na versão original e que serão utilizados como sinónimos ao longo do presente trabalho.

O número de actantes, isto é, de dependentes obrigatórios com os quais conta um verbo, constitui a valência quantitativa, que acaba por ser uma subclasse para a classificação e análise geral de unidades léxicas consideradas como portadoras de valência. No que concerne à valência quantitativa, pode-se falar de verbos zerovalentes (como *chover*), monovalentes (como *respirar*), bivalentes (como *comer*), etc. Não obstante, no contexto deste trabalho, é mais relevante falarmos em valência qualitativa, segundo a qual é determinado o comportamento morfossintático e semântico dos actantes que ocorrem num dado contexto e que será introduzida com maior detalhe ao longo desta pesquisa.

Não obstante, a já referida conceção multimodal também diz respeito à consideração de outras classes de palavras como potenciais portadores de valência, deixando para segundo plano a abordagem mais tradicional segundo a qual somente os verbos podem ser regedores de valência. Domínguez (2011, pp. 24-25) apresenta três conjeturas fundamentais que apareceram ao longo do desenvolvimento mais teórico da gramática valencial:

- Apenas os verbos podem ser contemplados como portadores de valência. Defendem esta hipótese autores como Brinkmann (1971), quem define a valência como sendo uma capacidade exclusiva do verbo.
- Outras categorias gramaticais, para além do verbo, podem reger um comportamento sintático-semântico determinado, isto é, podem contar com uma realização valencial. Schumacher (1986, p. 5) entende a valência como „die Fähigkeit von Subklassen der

morphologischen Klassen Verb, Adjektiv und Nomen, nach Zahl und Art bestimmte Ergänzungen zu sich zu nehmen.“ Outros autores, como Engel (2004), Sommerfeldt e Schreiber (1983) ou Zifonun (2003) também advogam esta formulação.

- Todas as unidades linguísticas podem ser portadoras de valência, incluindo as unidades multipalavra, como as expressões idiomáticas ou as colocações. Esta ideia é defendida por Ágel (2000), entre outros.

Embora fosse pertinente mencionar as diferentes hipóteses que existem para a consideração de determinadas classes de palavras como portadoras de valência, no presente trabalho não nos debruçaremos mais sobre esta questão. Portanto, defender-se-á a conjetura de os substantivos, entre outras categorias morfológicas ou gramaticais[7], poderem acarretar a realização de esquemas valenciais ou estruturas argumentais. Desde já, deve-se esclarecer o que é uma estrutura argumental:

> „Eine Argumentstruktur ist die semantische Repräsentation einer Proposition und besteht dabei aus genau einem Prädikat und der ihm zugeordneten Liste von Argumenten, wobei Prädikat und Argumente lexikalisch spezifiziert oder durch Variablen repräsentiert sind.“ (Engelberg, 2019, p. 15)

Nestes casos, deduz-se que o nome funciona como o predicado[8] de uma dada proposição semântica e que os argumentos[9] dependem dele. É essa hipótese a que configura o funcionamento da valência nominal (cf. Schierholz, 2001), segundo a qual os substantivos podem ter a capacidade de estruturar o seu contexto próximo com parâmetros de regência, facto pelo qual não se garante uma paráfrase bem-sucedida para uma ocasião linguística ou textual[10] qualquer (cf. Hölzner, 2007, p. 111). Seja como for, argumenta-se amiúde que a valência nominal se constitui como sendo um sistema *sui generis*:

> „Zum einen lassen sich längst nicht alle valenten Substantive auf Verben zurückzuführen (z.B. Straße nach Rom, Vorrat an Erdöl), zum anderen gibt es Ergänzungen beim Substantiv, zu denen Entsprechungen im verbalen Bereich fehlen (z.B. Genitivus partitivus), und schließlich lassen sich verbale Ergänzungen nicht systematisch und durch generelle Regeln

[7] É claro que os verbos podem funcionar como portadores de valência, pois são numerosos os estudos que se publicaram à volta desta consideração a partir de Teubert (1959). Podem-se citar, entre outros, Helbig e Schenkel (1969), Schumacher (1986), Curcio (1999) ou Engel e Savin (1983). Da mesma forma, nos últimos anos, proliferaram as investigações que consideram os adjetivos como portadores valenciais. Devem-se mencionar, nesta linha, as pesquisas de Sommerfeldt e Schreiber (1983a) ou de Matsekh-Ukrayinskyy (2015).

[8] Neste sentido, consulte-se Mel’čuk (2015, p. 9): “a predicate is a binding meaning, which has open slots for other meanings and binds them into a coherent complex meaning.”

[9] Para a delimitação dos argumentos e dos adjuntos, podem ser aplicados alguns testes linguísticos, entre os que se destacam o teste de substituição, de redução, de permutação, de extensão ou de introdução de cláusula relativa, entre outros (cf. Schierholz, 2001; Domínguez, 2011).

[10] A paráfrase que se menciona consiste frequentemente na substituição de um dado nome pelo verbo correspondente (*desejo* por *desejar*, por exemplo), verificando-se se se mantém uma estrutura em que os actantes possuem a mesma função linguística.

> beschreibbar in substantivische Ergänzungen überführen (z.B. *jemandem helfen/ Hilfe für jemanden*, aber: *jemandem danken/ Dank an jemanden*)." (Teubert, 1979a, p. 13)

Destarte, Domínguez (2011, p. 27) afirma que alguns complementos que aparecem na frase nominal podem ser específicos de uma dada subclasse morfológica dentro da categoria substantiva. É por isso que devemos fazer uma distinção entre as diferentes tipologias de nomes com as quais contamos, de modo a responder à seguinte questão: são todos os substantivos portadores de valência? Para resolver esta questão, Domínguez (2011, pp. 91-92) baseia-se em diferentes autores e aproxima-se das duas hipóteses mais relevantes. Por um lado, autores como Eisenberg (2001) defendem que apenas os substantivos derivados de verbos (deverbais) ou de adjetivos (deadjetivais) podem ser considerados como portadores valenciais, pois o comportamento actancial da base é mantido quando um determinado lexema é nominalizado. Por outro lado, investigadores como Teubert (1979, *vd. supra*) são da opinião de que todos os substantivos podem ser portadores de valência. Para tal fim, autores como Teubert (1979), Eroms (2000) ou Engel (2004) descrevem um inventário de actantes e circunstantes que podem aparecer em contexto nominal. Há que salientar que, tal como estuda Domínguez (2014b, p. 19), os inventários que os diferentes linguistas utilizam não elencam os mesmos actantes e circunstantes[11]. Este nível de heterogeneidade pode acarretar ainda complicações para a análise e descrição da valência nominal.

Do ponto de vista linguístico-gramatical, os substantivos podem ser, aliás, classificados como absolutos ou relativos. Helbig (1992, p. 123) assinala que os nomes absolutos carecem de valência semântica ou sintática. Por sua parte, os substantivos relativos remetem para predicados com um maior número de actantes, e deste grupo fazem parte os *nomina actionis* e os *nomina agentis*, sendo que apresentam no seu comportamento linguístico valência sintática e semântica. Os nomes selecionados para esta dissertação (*Bericht*, *Diskussion*, *Frage*) incluem-se no segundo grupo e trata-se, aliás, de substantivos deverbais (*Bericht* > *berichten*, *Diskussion* > *diskutieren*, *Frage* > *fragen*).

Antes de continuar com a descrição gramatical do comportamento valencial para a categoria nominal, parece relevante destacarmos os trabalhos que se realizaram no âmbito da lexicografia para contribuir à consolidação da valência nominal como sistema *sui generis* (cf. Teubert, 1979). Cumpre destacar o trabalho de Sommerfeldt e Schreiber (1983b), com o seu *Wörterbuch zur Valenz und Distribution der Substantive*, onde descrevem os esquemas argumentais de aproximadamente 750

[11] Em ocasiões, trata-se somente de divergências terminológicas entre os três autores. Não obstante, por vezes, o grau de heterogeneidade é maior e alguns autores reconhecem como sendo valenciais realizações formais que para outros não são (cf. Domínguez, 2014b, para, por exemplo, a discussão sobre a consideração dos adjetivos atributivos em Eroms).

verbetes, pois embora a maior parte dos substantivos aí recolhidos sejam derivados de verbos ou adjetivos, esta obra constitui-se como alicerce teórico para pesquisas ulteriores. Nos dias de hoje, também se deve considerar o trabalho realizado pela equipa de linguística e lexicografia da Universidade de Erlangen-Nürnberg, pois foi responsável pela criação de um dicionário *online* sobre a valência do substantivo em alemão[12] (cf. Schierholz, 2013), onde se presta especial atenção aos complementos prepositivos regidos por nomes.

No contexto bilingue, sobressai o trabalho de Kubczak e Constantino (1998), pois eles propõem a criação de um dicionário sintagmático com orientação valencial para o alemão e o francês, fornecendo informação gramatical sobre os portadores de valência e os respetivos argumentos de um ponto de vista contrastivo. Para o par de línguas polaco-alemão, Golonka (2002) descreve o comportamento sintático-semântico para padrões valenciais de verbos e nomes. Cumpre, aliás, destacar o trabalho de Bassola (2003), *Deutsch-ungarisches Wörterbuch zur Substantivvalenz*, que se caracteriza por ser um léxicon contrastivo alemão-húngaro baseando-se na valência nominal. Também não se deve esquecer a pesquisa de Stanescu (2008), quem se debruça sobre a descrição valencial de verbos, nomes e adjetivos em romeno e alemão. Por último, cumpre destacar o dicionário de valência substantiva desenvolvido por Domínguez (2011) para o espanhol e para o alemão, pois este volume revela-se como sendo um ponto de partida de referência para a descrição e análise de esquemas actanciais no nível frasal nominal.

Embora não se trate de uma abordagem valencial *stricto sensu*, pode-se referenciar ainda a Teoria Sentido-Texto (cf. Mel'čuk, 2015), pois para a criação de Dicionários Explicativos e Combinatórios (DEC) ressalta-se a formulação de padrões de regência como uma parte importante na microestrutura. Em consequência, estes padrões são representados mediante um nível de abstração que se leva a cabo com algumas variáveis. Nesta linha, Mel'čuk (2015, p. 114) afirma que devem levar-se em consideração vários aspetos: a dependência morfológica, a dependência lexical e sintagmática, e a dependência semântica, assim como todas as suas realizações superficiais.

A delimitação das unidades lexicais com as quais se trabalhará (*Bericht*, *Diskussion* e *Frage*), assim como a introdução aos parâmetros valenciais, servirão como fundamento teórico para a posterior análise metodológica. Contudo, deve-se ainda mencionar que não existe consenso na bibliografia teórica acerca de que realizações morfológicas que são indubitavelmente candidatas para a ativação de um determinado argumento. Tradicionalmente, falava-se em frases preposicionais como realização formal-morfológica predileta para a concretização de actantes (cf. Schierholz, 2001). Pesquisas posteriores,

[12] O dicionário em alemão pode ser consultado no seguinte link: http://www.erlanger-linguistik-online.uni-erlangen.de/projekte/erlangen-valency-patternbank.shtml.

entre as quais se pode mencionar a Kubczak e Schumacher (1998), Engel (2004), Bassola *et al.* (2004) ou Domínguez (2011), afirmam que os pronomes, os adjetivos ou as palavras compostas[13] em alemão (*Komposita*) podem igualmente ser actantes valenciais. No entanto, tal como acontece com as frases preposicionais (cf. *die Diskussion über die Reise* vs. *die Diskussion unter der Brücke*), deve-se considerar que não todos os adjetivos em posição atributiva atribuirão um significado argumental ao contexto frasal (cf. López, 2020, p. 13). Aliás, contempla-se a possibilidade de paráfrase, sendo que alguns argumentos podem ocorrer com mais de uma realização morfossintática e fornecendo mesmo assim um significado igual ou semelhante[14]. Neste sentido, Arias (2020, p. 25) esclarece esta idiossincrasia com as frases *die deutsch-amerikanische Debatte* (> *o debate germano-americano*) e *die Debatte zwischen Deutschen und Amerikanern* (> *o debate entre alemães e americanos*), uma vez que quer o adjetivo atributivo quer a frase preposicional desempenham o mesmo papel semântico, isto é, funcionam como agentes (ou seja, 'aqueles que debatem').

Estas realizações morfossintáticas dizem respeito à formulação e concretização do significado relacional, definido por Engel (1996, p. 226) a interação entre o predicado (neste caso, o substantivo central) e os complementos que ocorrem com ele. Tomando em consideração os objetivos específicos deste trabalho (*vd.* secção 2), o nosso foco centrar-se-á na delimitação do relator agente, pois será este o argumento com o qual se procederá na anotação semântica (semi)automática. Não obstante, cumpre recordar que existe um extenso inventário de argumentos[15] para o contexto frasal. Domínguez (2014b) elenca alguns dos principais complementos que podem aparecer no âmbito da frase nominal, entre os que se destacam o complemento sujeito (*Subjektivergänzung*), o complemento objeto (*Objektivergänzung*), o complemento prepositivo (*Präpositivergänzung*), o complemento adverbial (*Adverbialergänzung*), o complemento nominal (*Nominalergänzung*) ou o complemento verbal (*Verbativergänzung*). Relativamente a esta classificação, deve-se chamar a atenção para o facto de se tratar de uma categorização dos argumentos estritamente sintático-funcional, em oposição a outras tipologias (cf. Teubert, 1979) que misturam critérios semânticos e sintáticos.

Quanto ao complemento sujeito, foco de interesse especial no presente trabalho como se verá no capítulo 6, afirma-se que já Teubert (1979) e Engel (2004) falavam no argumento *Genitivus subiectivus*, segundo o qual se transmitia o conteúdo semântico do agente com diferentes realizações

[13] Relativamente aos compostos, Mel'čuk (2015, p. 121) destaca a produtividade dos lexemas construídos por composição morfológica em alemão, sendo que o primeiro elemento lexical do composto até pode funcionar como argumento.
[14] Oferecera-se mais informação a este respeito no capítulo 6 da presente dissertação, quando se aborde o tratamento específico de cada um dos substantivos apurados.
[15] Para uma descrição mais pormenorizada de cada um dos argumentos, consulte-se a bibliografia referenciada ou Teubert (1979) ou Engel (2004).

morfossintáticas, entre as quais se destacava uma frase em genitivo em língua alemã[16]. Domínguez (2014b, p. 22) esclarece que o complemento sujeito (*Subjektivergänzung*) é sobretudo realizado com uma frase em genitivo ou com uma frase preposicional que começa com a preposição *von* (> *de*). O teste para saber se se trata de um complemento sujeito consiste, segundo Domínguez (2014b, p. 22), na transformação da frase nominal questionada numa cláusula verbal a fim de se verificar se este complemento se converte em sujeito oracional (por exemplo: *die Diskussion zwischen Deutschen und Amerikanern* > *Deutsche und Amerikaner diskutieren*)[17].

Com este capítulo pretendia-se, em conclusão, providenciar uma visão geral sobre a gramática de valências e, nomeadamente, sobre a valência nominal. Destarte, este capítulo constitui o alicerce teórico sobre o qual se assentará a presente dissertação. Alguns dos elementos aqui mencionados acabarão por desempenhar um papel fundamental para a delimitação posterior da metodologia de pesquisa (*vd*. capítulo 5) e para a análise dos resultados atingidos (*vd*. capítulo 6).

[16] Na maior parte das línguas românicas, a correspondência com o genitivo alemão encontra-se na realização de frases preposicionais iniciadas com *de*. Desta forma, observa-se, no exemplo de Domínguez (2014b, p. 22) que o complemento sujeito da frase *die Klage* ***des Gemeinderates*** é traduzido com a preposição *de*: *a queixa* ***da Câmara Municipal***. Em consequência, no nível oracional, a realização seria *der Gemeinderat beklagt sich* ou, em português, *a Câmara Municipal queixa-se*, o que evidencia que *Gemeinderat* ou *Câmara Municipal* desempenha o papel semântico de agente.

[17] O comportamento é o mesmo em língua portuguesa: *a discussão entre alemães e americanos* > *alemães e americanos discutem*.

# 4

## Investigações relacionadas

O presente capítulo centra-se na descrição e apresentação de projetos e investigações realizadas no âmbito linguístico e lexicográfico que servem de orientação para a pesquisa aqui pretendida. Visa-se, em consequência, apresentar inicialmente diferentes pesquisas que contribuem, nos dias de hoje, para o aperfeiçoamento de recursos lexicográficos ou que representam uma tentativa para a melhor manipulação ou anotação de *corpora*. Neste sentido, serão abordados o projeto *FrameNet* e a abordagem de *Corpus Pattern Analysis* como referência para a anotação semântica (semi)automática que se persegue nesta dissertação. Depois, debruçar-nos-emos sobre os projetos PORTLEX, MultiGenera e MultiComb, cujo foco teórico reside na gramática de valências nominal, pois considera-se que podem supor um ponto de partida interessante para atingirmos alguns dos objetivos perseguidos.

### 4.1. *FrameNet*

O projeto *FrameNet*[18] foi inicialmente fundado pelo linguista C. Fillmore e continua a funcionar no Instituto de Ciências da Computação em Berkeley. Trata-se de uma base de dados lexicais para a língua inglesa cuja componente teórica reside na teoria semântica de *frames* e, num sentido mais amplo, na semântica lexical e na linguística cognitiva (Fillmore, 1985). Consoante o defendido por Fillmore (1985, p. 111), esta teoria constitui uma tentativa para descrever a organização mental das unidades lexicais numa dada cultura, isto é, uma iniciativa cujo objetivo principal é descrever estruturalmente um conjunto de situações (extra)linguísticas.

Fillmore (1982, p. 115) aponta para a relação que existe entre a semântica de *frames* e a gramática de valências, uma vez que ele próprio afirma que se trata de duas abordagens da língua cujo objetivo final consiste na descrição da combinatória lexical. No caso específico dos *frames*, deve-se assinalar que, através da sua consolidação, pretende-se descrever, do ponto de vista linguístico, uma série de eventos que aparecem associados, numa cultura, com uma unidade lexical. Desta forma, cumpre delimitar os participantes e os requisitos que articulam a situação típica, isto é, o *frame* (cf. Fillmore, 1977, 1982, 1985). A descrição do *frame* de uma unidade lexical fornece, em consequência,

[18] Para mais informações, consulte-se a página web do projeto: https://framenet.icsi.berkeley.edu/fndrupal/.

informação de carácter sintático-semântico, embora esteja fortemente ligada à cosmovisão de diferentes culturas.

No momento atual, tal como se pode ver na base de dados *FrameNet*, já foram descritas 13685 unidades lexicais e delimitaram-se 1124 *frames*. Deve-se ainda chamar a atenção para o facto que serem nomes as unidades lexicais com maior representação, embora os *frames* costumem aparecer relacionados com verbos. A anotação proporciona informação de tipo semântico-relacional e classifica os elementos como sendo essenciais ou periféricos, entre outras categorias, para a unidade lexical selecionada.

Na medida em que na presente dissertação, a anotação semântica (semi)automática pretendida será desenvolvida para a língua alemã, nomeadamente a existência do projeto SALSA[19], da Universidade de Saarland, no âmbito do qual se realizou uma anotação manual de lexemas verbais atualmente digitalizada. Para tal objetivo, a equipa responsável pelo desenvolvimento deste trabalho lançou mão da teoria semântica de *frames* sempre que foi possível (cf. Baker, 2012). A anotação sintática baseia-se no *corpus* de textos jornalísticos em alemão TIGER, partindo do qual Marek (2009) pretendeu realizar uma extensão que consistia na extração de dados linguísticos não só com consultas *query* sintáticas, senão que também visava possibilitar a retirada de informação do *corpus* com *queries* adaptadas aos *frames* já definidos.

### 4.2. *Corpus Pattern Analysis* (CPA)

A abordagem do *Corpus Pattern Analysis* (doravante, CPA) baseia-se na Teoria de Normas e Explorações (*Theory of Norms and Exploitations*; doravante, TNE) desenvolvida pelo linguista P. Hanks (cf. 2004, 2013), cujo objetivo consiste em identificar, através de estratégias de manipulação de corpora, realizações superficiais e comportamentos semânticos para determinadas unidades lexicais. Trata-se, neste sentido, de uma técnica de tratamento lexical que visa conhecer o léxico de uma língua ao descrever os padrões em que uma dada unidade lexical ocorre.

Para a descrição lexical mediante padrões, Hanks (2013, p. 177) insiste na importância de delimitar *sets* lexicais cuja configuração interna inclui diferentes tipos semânticos. Desta forma, os *sets* lexicais podem estar formados por um conjunto ilimitado de unidades lexicais correspondentes a um traço semântico específico, por exemplo, [[humano]]. Não obstante, o *set* [[humano]] possibilita a sua análise interna, de forma que se distinguem tipos semânticos como [[humano instituição]], o que pode

[19] Para mais informações, consulte-se https://www.coli.uni-saarland.de/projects/salsa/.

acarretar diferenças de sentido para o conjunto co(n)textual. Do ponto de vista da teoria CPA, a tipologia semântica de uma determinada unidade lexical pode ser coagida pelo contexto (cf. Hanks 2013, p. 219).

Um dos resultados desta técnica de tratamento lexical é o *Pattern Dictionary of English Verbs* (PDEV)[20], cujo processo de automatização atual graças a técnicas de PLN contribui para um desenvolvimento mais rápido da descrição lexical verbal (cf. Hanks, 2013). Atualmente, o dicionário fundamenta-se na ontologia CPA, pois o autor aponta para a importância de organização semântica, de modo a atingir a análise de padrões lexicais pretendida.

Esta ideia também é apresentada e defendida por Renau e Nazar (2016a), pois introduzem, a partir de análises de corpora seguindo a aproximação de Hanks (2004, 2013), padrões lexicais com verbos em espanhol. Renau e Nazar (2016a, p. 826) salientam a importância da taxonomia ontológica para conseguirem uma análise de padrões lexicais e de estruturas argumentais, pois a possibilidade de filtrar e mapear os tipos semânticos com técnicas de PLN facilita a aproximação metodológica.

Conclui-se, portanto, que a análise de padrões lexicais também reflete, até certo ponto, a presença de argumentos em esquemas valenciais, uma vez que a dependência semântica é um aspeto importante e comum às duas abordagens teóricas. Destarte, e seguindo as teses subjacentes à técnica CPA, acabaria-se por conseguir, no caso da anotação semântica de estruturas argumentais nominais esquemas como os seguintes: *discussão de* [[humano]] ou *discussão de* [[imaterial]]. Assim sendo, atingir-se-ia igualmente o objetivo da extração de dados linguísticos já classificados de acordo com categorias semânticas.

### 4.3. Os projetos PORTLEX, MultiGenera e MultiComb

No âmbito da valência nominal de orientação multilíngue, tem-se avançado de forma significativa ao longo das últimas décadas. O projeto PORTLEX[21] é um claro exemplo disto, pois trata-se, fundamentalmente, de uma "lexicographical online tool that compiles multilingual data on the valency of the noun phrase in German, Galician, Spanish, Italian, and French" (Domínguez & Valcárcel, 2019, p. 135). A teoria apresentada na secção 3.2., assim como o volume de Domínguez (2011), subjazem ao desenvolvimento desta ferramenta e são considerados como os fundamentos teóricos do projeto aqui referido.

Deve-se sublinhar que o projeto PORTLEX assenta as bases para a futura criação de dicionários valenciais ao nível da frase nominal, graças à análise e anotação detalhada que se realiza nesta

[20] Consulte-se aqui: https://pdev.org.uk/.
[21] A ferramenta online pode ser consultada no seguinte link: http://portlex.usc.gal/diccionario/

ferramenta para uma série de nomes. Aliás, e de acordo com Domínguez e Valcárcel (2019), não se consideram apenas substantivos deverbais ou deadjetivais, senão que também se incluem outros sem correspondentes no contexto verbal ou adjetival, como é o caso dos nomes *problema* ou *gana*. Para além desta consideração mais abrangente, introduzem-se na plataforma PORTLEX realizações formais do ponto de vista morfossintático que não estão recolhidas no dicionário de Sommerfeldt e Schreiber (1983b), como é o caso dos adjetivos, das palavras compostas em alemão ou das sequências de dois nomes, $N_1N_2$ (cf. Valcárcel, 2016, para o caso do francês), tratando-se, portanto, de estruturas desprovidas de preposição.

Assinala-se que o PORTLEX foi, conforme o defendido por Mirazo (2016, pp. 94-95), concebido para a resolução de problemas interlinguísticos e de intermediação entre as línguas que atualmente estão aí incluídas, prestando-se atenção especial à construção da frase nominal. Segundo Dušek (2013, p. 28), verifica-se que os dicionários de valência costumam desempenhar um papel crucial na produção linguística em língua estrangeira, pois a maior parte dos problemas de interferências entre a língua de origem e a de destino aparecem na codificação (neste sentido, cf. Iriarte, 2004).

Uma novidade incluída no PORTLEX e que o diferencia de outros dicionários anteriores consiste na possibilidade de interação dos utentes, pois estes podem manipular dados (por exemplo, corrigir ou introduzir dados) e, após uma revisão da equipa lexicográfica, a informação pode acabar por ser publicada e disponibilizada online. Trata-se, neste sentido, de uma obra de consulta de carácter colaborativo. A estrutura modular que permite uma comparação em tempo real entre as estruturas sintagmáticas de várias das línguas aí recolhidas faz com que o dicionário ganhe em riqueza do ponto de vista da análise multilíngue, pois possibilitam-se pesquisas quer monolingues quer multilingues para várias línguas de partida e de chegada (cf. Domínguez & Valcárcel, 2019).

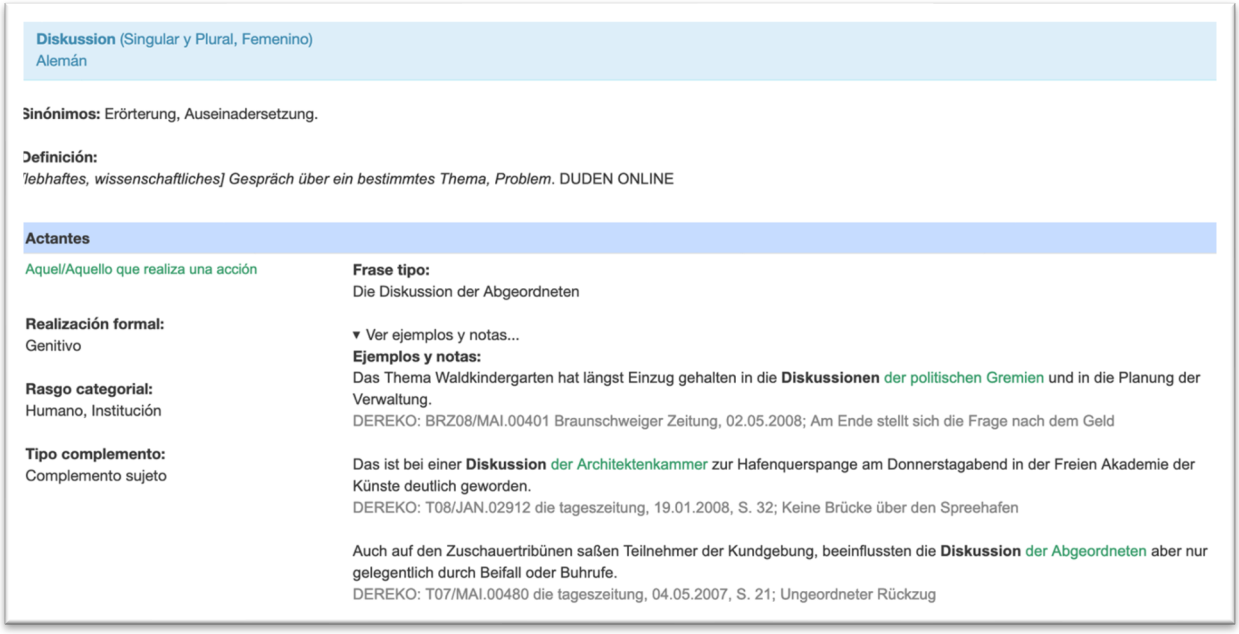
No que diz respeito aos fundamentos do PORTLEX mais pertinentes para os objetivos desta dissertação, cumpre mencionar a conformação da microestrutura (*vd.* figura 2), pois para além da paráfrase de significado e da parte paradigmática onde se incluem sinónimos, a compreensão da informação sintagmática acaba por tornar-se essencial para um uso adequado desta ferramenta. Destarte, para um dado actante (na figura 2, 'aquele/aquilo que realiza a ação')[22], apresentam-se as

[22] No manual de utilizador do dicionário *PORTLEX*, disponível na sua página web, elencam-se os diferentes actantes sintático-semânticos que podem encontrar-se no dicionário de modo a possibilitar uma pesquisa avançada. Ressalta-se que se empregam expressões de fácil compreensão caso os utentes não sejam expertos em linguística: "aquele/aquilo que realiza uma ação", "aquele/aquilo que é afetado", "aquele/aquilo que existe ou que é", etc.

potenciais realizações formais acompanhadas com uma anotação semântica (na figura 2, observa-se que o complemento sujeito para *Diskussion* tem que ser [+humano] ou [+instituição]).

**Figura 2**

*Captura de ecrã do artigo lexicográfico* Diskussion *em PORTLEX*




*Nota*. Na microestrutura, é especialmente importante a inclusão de informação relativa à estrutura argumental (principal objetivo do dicionário citado), pois o dicionário PORTLEX oferece dados lexicográficos que dizem respeito ao comportamento semântico e sintático ou formal do lema consultado.

Tal como indicado na secção 3.2. do presente trabalho, a frase em genitivo é uma das possibilidades de realização morfossintática para o complemento sujeito deste substantivo, entre outros. A importância de incluir esse tipo de informação sintática e semântica remete para a afirmação de Domínguez *et al*. (2019, p. 35), segundo quem "a valency dictionary should provide syntactic and semantic information that helps its users to improve the linguistic production in a foreign language". Neste sentido, a inclusão de exemplos extraídos de *corpora* constitui uma parte essencial da zona de coocorrência léxica do PORTLEX (cf. Domínguez & Valcárcel, 2019, p. 147), sendo que se garante o oferecimento de construções de língua natural onde ocorrem as estruturas sob análise e descrição.

A anotação semântica do significado categorial, para além da descrição de estruturas argumentais, representa o contributo mais importante do PORTLEX para a análise que será realizada nesta dissertação. Deste modo, verifica-se que os actantes semânticos, ou a paráfrase dos papéis semânticos (agente, paciente, tema...), aparecem descritos acompanhados pelos traços [+humano] e [+instituição], de forma que ajuda os utentes, entre os que se destacam estudantes de línguas estrangeiras, a produzir novas sequências.

O desenvolvimento e implementação do projeto PORTLEX requereu um recurso constante a *corpora* para a introdução de exemplos de língua real. Foi ao longo desse processo que se revelou necessário realizar uma revisão pormenorizada de *corpora* já existentes:

> "The work done for the PORTLEX dictionary [...] has highlighted the limitations of corpus-driven methods in achieving the objective of a lexicographical project, at least from a dependency-valency perspective [...]. Corpora do not contain examples of all realizations of all the realizations of the different actants or of their combinations. On the other hand, the examples found in corpora sometimes do not meet the intelligibility or conciseness requirements of a dictionary." (Domínguez *et al.*, 2018, p. 847)

A partir das experiências no dicionário PORTLEX, a equipa de trabalho decidiu desenvolver recursos para geração automática de língua, visando a geração automática de exemplos para dicionários de valências. As ferramentas cujo objetivo consiste na geração de estruturas foram desenvolvidas no âmbito dos projetos MultiGenera e MultiComb, que ultrapassam os limites da gramática valencial (cf. Domínguez, 2022). Estes projetos baseiam-se, aliás, noutras teorias, entre as que se deve ressaltar a Teoria Sentido-Texto (cf. Mel'čuk, 2015), e graças às quais se puderam atingir objetivos salientáveis e resultados favoráveis em Geração de Língua Natural (GLN). Quanto à GLN, deve-se sublinhar que é uma subdisciplina importante do PLN, sendo que é aquele processo que acaba por possibilitar a satisfação do objetivo comunicativo do emissor (cf. Bernardos, 2007). Destarte, tal como se afirma em Bernardos (2007, p. 106), cumpre lançar mão de algoritmos e ferramentas que permitam escolher que elementos serão utilizados e quais não para uma comunicação bem-sucedida em língua natural gerada de forma automática.

Estes princípios, assim como a teoria linguística apresentada em § 3, subjazem ao desenvolvimento dos projetos MultiGenera e MultiComb, considerando que se visa fundamentalmente gerar estruturas argumentais nominais tomando informação semântica de bases de dados léxicas, em concreto de WordNet[23]. Assim, e como experiência preambular, Valcárcel e Domínguez (2016) dirigiram um projeto-piloto onde falantes ocasionais avaliaram a correção de frases geradas automaticamente com o substantivo *muerte*. Embora a validade empírica absoluta deste projeto seja questionada pelos próprios investigadores devido ao facto de tratar-se de um pequeno inquérito que não possibilita uma retirada objetiva de conclusões definitivas, deve-se destacar que serviu como esboço experimental para a

[23] Consulte-se aqui: https://wordnet.princeton.edu/.

confirmação de algumas hipóteses e ideias que acabaram por tentar verificar-se de forma definitiva com a execução dos projetos posteriores, MultiGenera e MultiComb.

Para o correto desenvolvimento dos geradores piloto, foram concebidas várias ferramentas (*MultiTools*) de análise e extração de dados concebidas como parte dos projetos *supra* mencionados. Deste modo, a metodologia aplicada em MultiGenera e MultiComb assenta em três alicerces: "i) the automatic extraction of data from NLP resources, ii) the analysis of corpora, co-occurrence databases and wordnets, iii) as well as the outcoming evaluation produced by both generators" (Domínguez *et al.*, 2019, 53). A seguir, serão apresentadas algumas das ferramentas que são essenciais ao desenvolvimento do presente trabalho: trata-se, fundamentalmente, da ontologia lexical criada[24] e do gerador *Xera*[25].

Em primeiro lugar, para a criação da ontologia lexical desenvolvida e que se utilizará nesta investigação, foi necessário o recurso assíduo às ferramentas *Lematiza*, *Combina* e *Flexiona*[26]. No que diz respeito a *Lematiza*, trata-se de um recurso que possibilita a lematização automática de dados extraídos de Sketch Engine e incluídos num esquema actancial determinado, de modo que podem ser observados posteriormente ao aparecerem ligados às diferentes ontologias manipuladas em WordNet. Para além disso, apresentam-se os *synsets* com os quais uma unidade lexical está associada, de forma que se pode aceder diretamente aos dados recolhidos numa dada categoria (cf. Domínguez *et al.*, 2019). Em consequência, WordNet revelou-se primordial para a criação e o desenho da ontologia lexical própria, onde aparecem registados os traços semântico-categoriais dos diferentes protótipos lexicais, considerando-se sempre os espaços funcionais valenciais concretos em cada caso. Consoante Domínguez (2021, p. 31), lembra-se que:

> „Als Prototype fassen wir typische bzw. repräsentative Instanzen für eine konkrete Argumentstruktur oder Slotbesetzung auf. Dabei handelt es sich nicht um semantische abstrakte Konzepte, da der Prototypenbegriff mit der syntaktischen Argumentstruktur eines konkreten nominalen Valenzträgers sowie mit der Aktualisierung einer konkreten Bedeutungslesart zusammenhängt [...]. Ihre Typikalität sowie Repräsentativität lässt sich auf die Interaktion valenzfundierter und frequenzbezogener Parameter stützen [...]".

[24] Consulte-se aqui: http://portlex.usc.gal/ontologia/.
[25] Para consultas, veja-se a *website*: http://portlex.usc.gal/develop/xera.php?.
[26] Estas ferramentas são conhecidas, no âmbito dos projetos MultiGenera e MultiComb, como MultiTools e estão disponíveis no seguinte link: http://portlex.usc.gal/combinatoria/.

Domínguez (2022, p. 173) chama a atenção para o facto de não interessar tanto ver se as unidades lexicais espanholas *cabeza* ou *pelo* são protótipos gerais para a categoria {parte do corpo humano}, senão que o foco centra-se na adaptação de critérios funcionais e de frequência para possibilitar conclusões como que *cabeza* é um protótipo dessa classe semântica na estrutura argumental *el dolor de* + {parte do corpo humano}, enquanto *pelo* pode aparecer como protótipo no esquema *el color de* + {parte do corpo humano}.

Uma vez esclarecido o funcionamento de criação de protótipos para a ontologia, deve-se esclarecer como se consegue a fase de expansão lexical. Para tal objetivo, recorre-se à ferramenta *Combina*[27], tomando como apoio a informação fornecida pelas ontologias inseridas no WordNet[28]. *Combina* realiza uma consulta de APIs previamente desenhadas pela equipa informática e combina todas as unidades lexicais recolhidas nas ontologias de WordNet (cf. Gómez & Solla, 2020). Graças a esta ferramenta, consegue-se fazer uma pesquisa que nos oferece, como resultado, o vocabulário existente para as classes semânticas determinadas, para além de uma lista de *synsets* (cf. Domínguez *et al.*, 2019).

A lista de vocabulário conseguida através da ferramenta *Combina* pode ser descarregada para o ser flexionada no recurso *Flexiona*[29]. Posteriormente, que a lista de unidades lexicais é anotada de forma sintática e morfológica para a posterior criação dos pacotes lexicais (seguindo a nomenclatura da equipa encarregada dos projetos) que são representantes da ontologia lexical *supra* mencionada. No que concerne aos pacotes lexicais, deve-se explicar que:

> "A lexical package [...] describes a set of related lexical units that, although not interchangeable, have a similar paradigmatic relationship. As an extremely simple example, despite their different meaning, distinct parts of the human body can be used in similar structures [...]". (Domínguez *et al.*, 2021, p. 277).

Deve-se ainda sublinhar que, consoante o defendido por Rodríguez *et al.* (1998, p. 132) não existe um consenso na definição de ontologias, senão que o único fator comum nos diferentes recursos categorizados como ontologias consiste na aproximação à organização do conhecimento. Deste modo, possibilita-se um futuro reaproveitamento da informação para outras aplicações. Para além disso, Rodríguez *et al.* (1998, p. 132) assinalam que as ontologias diferem na delimitação do seu conteúdo e

[27] A ferramenta *Combina* está disponível sob a seguinte hiperligação: http://portlex.usc.gal/develop/combina.php

[28] Para o funcionamento de *Combina*, é essencial operar com os links que remetem para a ontologia, como por exemplo: http://portlex.usc.gal/develop/de/api/?ontology=sumo&category=Human&subcategories=on, pois a API precisa de ser copiada na ferramenta para a realização da consulta.

[29] Este recurso foi disponibilizado aqui: http://portlex.usc.gal/develop/flexiona.php.

no tratamento do mesmo, sendo que algumas só têm o objetivo de servir para pesquisas terminológicas e/ou terminográficas. A ontologia lexical que utilizaremos para o presente projeto foi criada, como *supra* mencionado, com o propósito de descrever semanticamente *slots* funcionais do ponto de vista valencial, o que faz com que se considere apropriada para os objetivos aqui perseguidos. Deste modo, Domínguez (2020, p. 74) sustenta que o estabelecimento dos protótipos lexicais e, consequentemente, dos pacotes lexicais, é fundamental para o processamento e geração automática de dados das ferramentas em questão.

Com efeito, as estruturas organizadas em ontologias, permitem-nos falar numa composição em vários níveis, tal como apontam Rodríguez *et al.* (1998) ou González e Rigau (2013):

- Entidade de primeiro nível (*1st order entity*): correspondem, normalmente, a entidades gerais e percetíveis através dos sentidos por existirem no mundo extralinguístico.
- Entidade de segundo nível (*2nd order entity*): dizem respeito a estados ou situações que costumam ser apercebidas como independentes.
- Entidade de terceiro nível (*3rd order entity*): as entidades deste nível são normalmente definidas como ideias ou conceitos mentais que podem ser corroborados ou negados.

Defende-se, a partir do exposto, que a ontologia lexical de Domínguez *et al.* foi elaborada de forma *bottom-up*, partindo dos dados de frequência e de consultas CQL realizados em Sketch Engine. Portanto, a ontologia apresenta-se também num estágio de constituição constante, uma vez que a consulta de novas estruturas argumentais ou novas unidades lexicais acaba por alargar constantemente o seu tamanho. Para a criação da ontologia, utilizou-se o inventário de traços categoriais de Engel (2004). Para uma análise pormenorizada de classes semânticas, Engel (2004, p. 185) defendia que o significado de uma determinada unidade lexical era essencial pelo facto de ser "inerente". Aliás, fala-se no significado categorial como sendo ponto de partida indiscutível para uma descrição semântica de quaisquer unidades lexicais (cf. Engel, 2004, p. 188). A ontologia baseia-se, portanto, neste inventário, que foi ampliado com o recurso às anotações existentes nas ontologias incluídas em WordNet (cf. Domínguez *et al.*, 2019).

No que diz respeito à construção dos pacotes lexicais manipulados em conjunção com a ontologia lexical de Domínguez *et al.* (2021), deve-se considerar que o recurso a WordNet permite afinar a descrição semântica de modo que se providencia informação organizada num maior número de níveis, pelo que se atinge uma maior granularidade. Gross (2013) aponta para a existência de várias classes semânticas dependendo dos padrões sintático-semânticos. Para além disso, Gross (2013, p. 120) chama

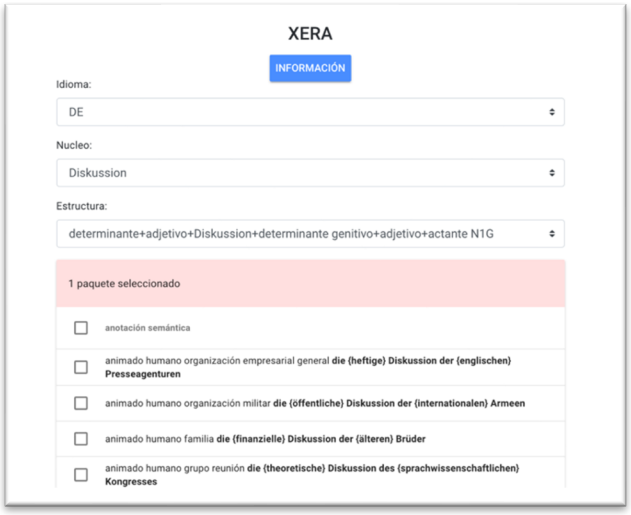
a atenção para a dificuldade de classificar algumas unidades lexicais do ponto de vista semântico, pois alguns nomes de organizações ou instituições, por exemplo, apresentam um grau elevado de ambiguidade, uma vez que se podem referir a lugares (locativos) ou à instituição *per se* (agente). Isto acontece com substantivos como *escola* ou *empresa* e precisa-se de um contexto maior para clarificar o significado.

A ontologia lexical com que se trabalhará contempla especialmente o plano paradigmático e as relações de sinonímia, hiponímia, hiperonímia, entre outras, que existem entre as unidades lexicais aí incluídas. Não obstante, no âmbito dos projetos MultiGenera e MultiComb presta-se atenção também ao plano sintagmático, pois ferramentas como *Xera*, *Combinatoria* ou *CombiContext* têm por objetivo a geração automática de estruturas linguísticas a diferentes níveis que serão abordados mais pormenorizadamente a seguir.

O primeiro dos protótipos mencionados, *Xera*[30], permite a geração automática de estruturas monoargumentais ao nível da frase nominal (cf. Domínguez, 2022). Escolhendo um esquema actancial para um determinado substantivo que funciona como núcleo ou predicado, são gerados exemplos ligados às classes semânticas escolhidas (*vd.* figura 3 *infra*).

**Figura 3**

*Captura de ecrã de* Xera




*Nota*. Esta captura de ecrã foi realizada em *Xera*. A ferramenta permite escolher a língua de trabalho (neste caso, alemão), o núcleo nominal (neste caso, *Diskussion*), a estrutura argumental e o(s) pacote(s) semântico(s). A determinação de toda esta informação estabelece o ponto de partida para a geração automática de frases monoargumentais.

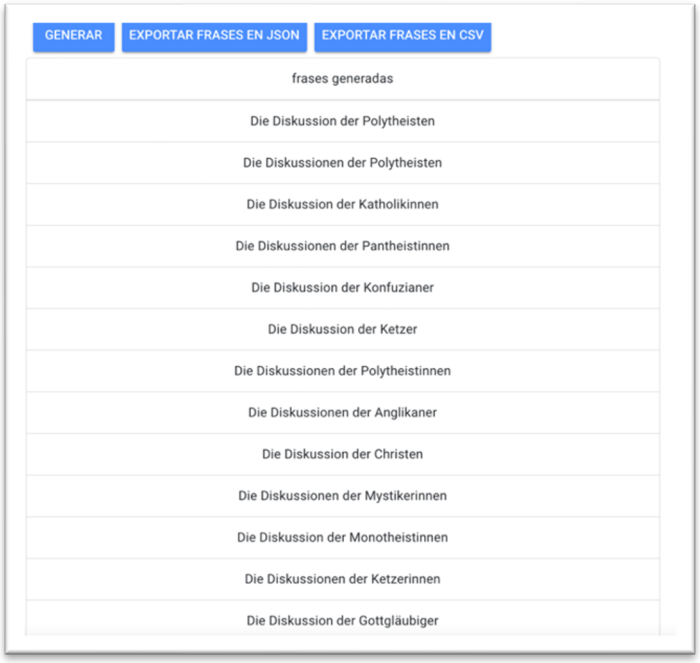
Em geral, estes exemplos são considerados como gramaticais e válidos para serem incluídos em recursos lexicográficos. Para ilustrar o referido, uma seleção da estrutura argumental "determinante

[30] Consulte-se aqui: http://portlex.usc.gal/combinatoria/usuario.

+ adjetivo (opcional) + *Diskussion* + determinante genitivo + adjetivo (opcional) + actante N1G[31]" em combinação com a classe semântica {crença religiosa}, fornece-nos exemplos automáticos como *die Diskussion der Agnostiker* ou *die Diskussionen der Atheistinnen*, entre outros que se apresentam na figura 4 *infra* e que podem ser exportados em formatos JSON ou CSV para manipulações posteriores.

**Figura 4**

*Captura de ecrã de* Xera



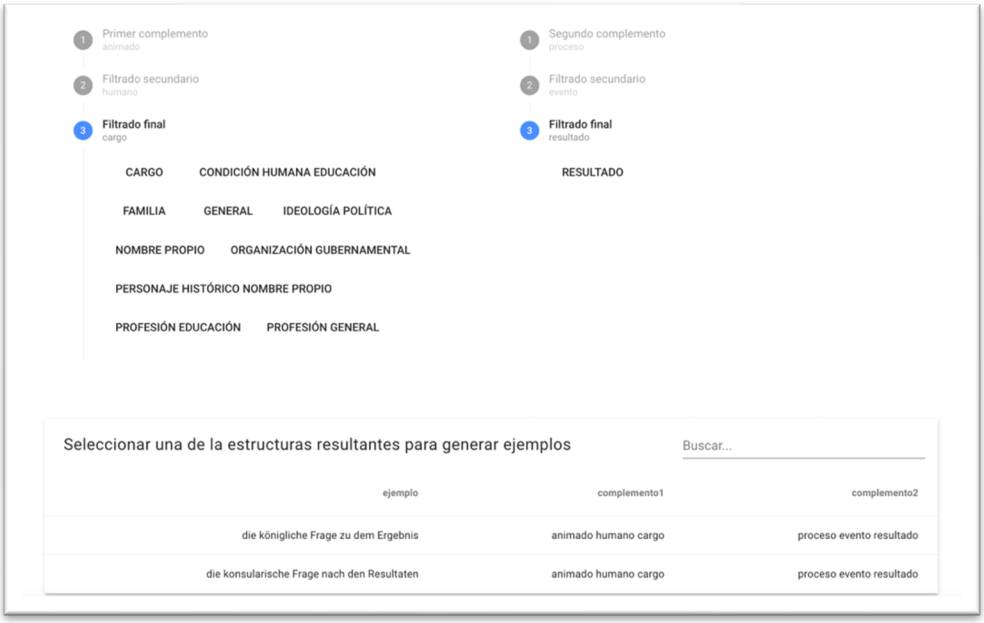

*Nota*. Esta captura de ecrã foi realizada em *Xera*. A ferramenta permite exportar as frases geradas automaticamente em formato JSON ou CSV, o que facilita a posterior manipulação dos dados linguísticos proporcionados.

O segundo dos protótipos, *Combinatoria*[32], visa gerar automaticamente estruturas biargumentais (cf. Domínguez, 2022). Para tal fim, a ferramenta permite aos utentes selecionar as classes semânticas dos dois actantes que fazem parte da estrutura, para além da organização que os complementos têm no contexto frasal. A seguir, geram-se automaticamente frases que cumprem os requisitos pré-selecionados, tal como se pode observar na figura 5, estabelecendo-se uma abordagem semântico-conceitual e não de carácter formal.

[31] Abreviatura utilizada para fazer referência à frase nominal para o primeiro argumento realizada em genitivo.
[32] Disponível no seguinte link: http://portlex.usc.gal/combinatoria/.

**Figura 5**

*Captura de ecrã de* Combinatoria




*Nota*. Esta captura de ecrã foi realizada em *Combinatoria*. A ferramenta possibilita gerar, de forma automática, padrões com dois argumentos a partir de uma seleção de traços semânticos.

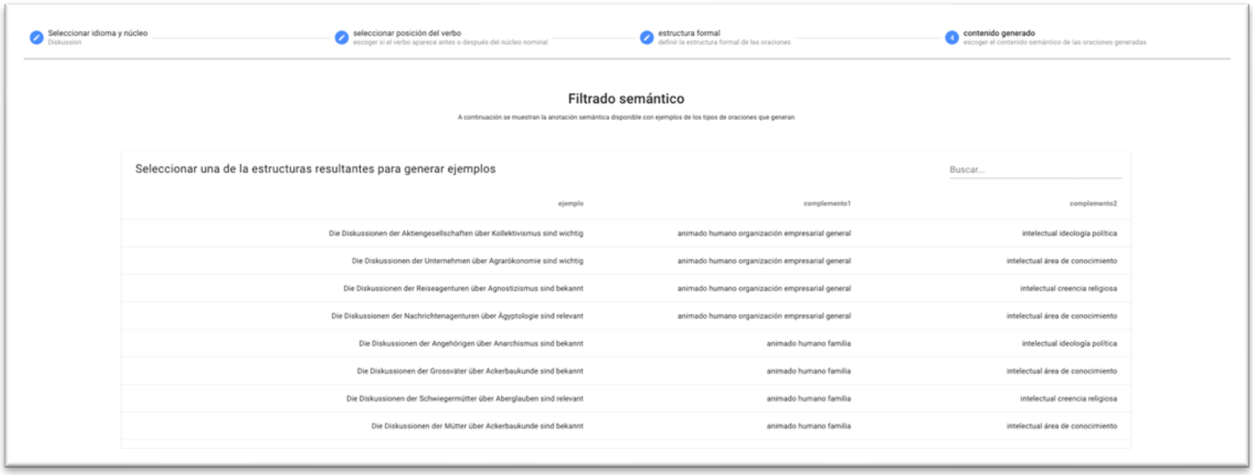
O terceiro protótipo, *CombiContext*[33], providencia um contexto oracional às frases biargumentais geradas com as outras ferramentas (cf. Domínguez, 2022). Desta forma, conseguem-se integrar as estruturas frasais nominais em contextos discursivos. Fornece-se, portanto, informação relativa ao co(n)texto onde podem ocorrer as frases nominais com uma estrutura semântico-funcional predeterminada (*vd*. figura 6 *infra*).

Em suma, pode-se deduzir que os contributos decorrentes tanto do dicionário PORTLEX como dos projetos MultiGenera e MultiComb desempenharão um papel essencial na presente dissertação. Em consequência, em § 5, centrar-se-á o foco naqueles aspetos que se consideram cruciais para o a abordagem metodológica aqui pretendida, sendo que alguns dos resultados dos projetos *supra* citados serão fundamentais, especialmente pelo facto de partilharem com esta pesquisa uma parte importante do seu enquadramento teórico (cf. Domínguez, 2011).

[33] Pode ser consultado aqui: http://portlex.usc.gal/combinatoria/verbal.

**Figura 6**

*Captura de ecrã de* CombiContext




*Nota*. Esta captura de ecrã foi realizada em *CombiContext*. A ferramenta integra as frases nominais geradas automaticamente em *Xera* e *Combinatoria* num quadro oracional, permitindo escolher filtros semânticos e posições dos argumentos a respeito do verbo na oração.

# 5

## Enquadramento metodológico

Para atingirmos os objetivos anteriormente elencados (*vd*. capítulo 2), cumpre, neste ponto, estabelecermos uma metodologia adequadamente delimitada que nos permita centrar o foco no nosso objeto de estudo, isto é, na anotação semântica automática de estruturas argumentais no nível frasal em alemão. Este quadro metodológico (*vd*. secção 5.2.) condicionará a posterior análise dos resultados, assim como a retirada de conclusões relativas às nossas hipóteses atuais de trabalho (*vd*. secção 5.1.). Para tal fim, neste capítulo apresentar-se-ão as perguntas de investigação que articulam o presente trabalho, assim como a estratégia metodológica selecionada, sendo que desempenharão um papel essencial alguns dos problemas, do ponto de vista da aplicação da metodologia, que teremos que defrontar ao longo da presente pesquisa.

### 5.1. Perguntas de investigação

Na literatura científica, algumas autoras que se debruçaram sobre a gramática de valências de um ponto de vista menos tradicional pela incorporação de metodologia proveniente da área de PLN, entre as quais cabe mencionar Domínguez *et al.* (2018) e López (2020), apontam para o facto de os *corpora* linguísticos não disporem de anotação semântica. A falta de anotação quer semântico-relacional quer semântico-ontológica[34] de entidades linguísticas em *corpora* acarreta uma série de dificuldades para a extração de estruturas argumentais. Em consequência, a presente dissertação pretende contribuir para o tratamento semântico da informação da qual dispomos em *corpora*, pois assim facilitar-se-ia a posterior extração de informação para possíveis aplicações lexicográficas, por exemplo. Neste sentido, e de modo a aumentar o grau de coerência da presente pesquisa, formulam-se as seguintes perguntas de investigação:

- Consegue-se automatizar, pelo menos parcialmente, através da criação de um *script* informático *ad hoc*, um sistema para anotarmos semanticamente unidades lexicais em *corpora* linguísticos? Se a resposta for positiva, em que medida é plausível pô-lo em

[34] Fala-se em anotação semântico-relacional para nos referirmos aos papeis semânticos que uma unidade lexical desempenha num contexto em relação ao predicado, como por exemplo, agente, paciente, etc. Por sua parte, a anotação semântico-ontológica diz respeito às categorias semânticas intrínsecas a uma unidade lexical, também no sentido da teoria CPA (cf. Hanks, 2013), como por exemplo, humano, situação, imaterial, etc. Para esta distinção, veja-se também Domínguez *et al.* (2021).

prática? Para além disso, quais são as dificuldades e desafios que esta abordagem comporta?

Para tentarmos procurar respostas às perguntas aqui apresentadas, trabalha-se com a hipótese de que tal como as consultas CQL morfossintáticas facilitam a manipulação de dados a nível formal em *corpora* linguísticos como o Sketch Engine (Jakubíček *et al.*, 2010), supõe-se que uma anotação semântica automática e estruturada adequadamente poderia conduzir a resultados positivos no que concerne à extração de informação semântica.

A potencial efetividade da interface de anotação semântica levantará questões que dizem respeito à componente mais linguística do trabalho, sendo que parece relevante refletir sobre o seguinte assunto:

- Sendo a polissemia uma realidade muito habitual na língua natural, interessa descobrir se através de uma automatização de anotação e extração de informação a partir de *corpora* se pode contribuir para a desambiguação semântica.

A hipótese subjacente à questão agora abordada tem a ver com o facto de alguns significados serem ativados graças a determinadas realizações formais no nível da frase nominal, neste caso. Iriarte (2001) defende que existem mecanismos linguísticos cuja realização superficial facilita a desambiguação de casos de polissemia. Neste sentido, Iriarte (2001) remete para o trabalho de Calderón (1994, pp. 54-55):

> "normalmente, una distinción de significados lleva aparejada una diferenciación formal, es decir, que a acepciones distintas de una misma palabra suele corresponderle estructuras formales también diferentes [...]. Es decir, las diferentes acepciones de una palabra no radican exclusivamente en dicha palabra, sino en ella más otros elementos de su entorno [...]".

Todavia, na presente dissertação, serão analisados casos em que a realização puramente formal é insuficiente para a conclusão de alguns estudos e daí surge a necessidade de anotação semântica, uma vez que o plano semântico é a única possibilidade de desambiguação quando duas estruturas são exatamente idênticas superficialmente (cf. *die Frage der Teilnehmer* vs. *die Frage der Armut*). Em síntese, as perguntas de investigação apresentadas neste subcapítulo, assim como algumas das hipóteses de trabalho, permitirão estruturar o quadro metodológico com que se operará a seguir. Desta forma, introduzir-se-ão na seguinte secção (5.2.) aspetos cruciais para a obtenção e extração de

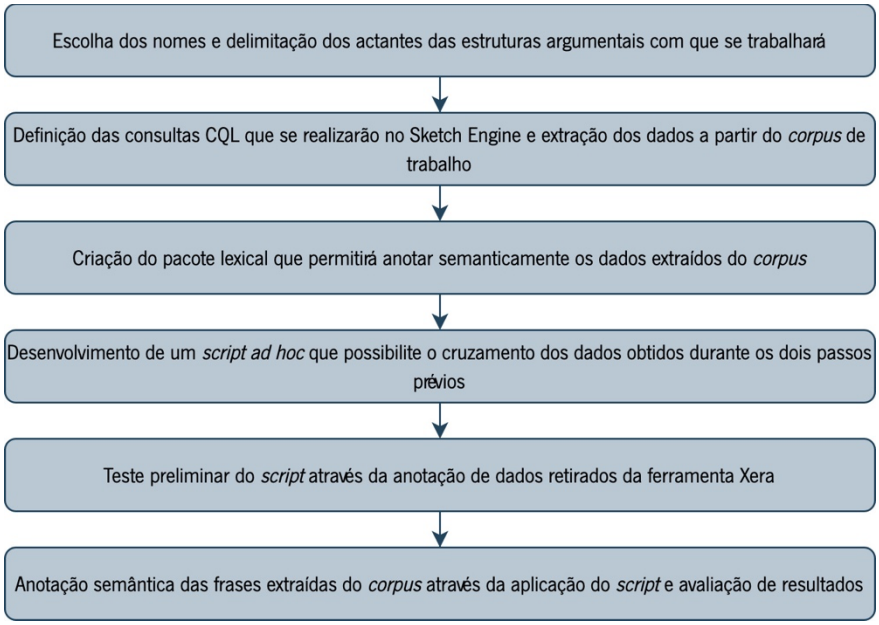
dados linguísticos de *corpora*, assim como para a posterior aplicação informática e finalmente para a avaliação dos resultados atingidos.

## 5.2. Metodologia

No que diz respeito à metodologia aplicada, deve-se considerar que se definiram inicialmente diferentes fases de trabalho que serão explicadas agora de forma mais pormenorizada. Na figura 7, pode-se observar o elenco de passos pré-definidos que nos permitiram estruturar de forma clara a presente pesquisa de modo a conseguirmos uma avaliação posterior de resultados.

**Figura 7**

*Fases de trabalho metodológico*




A escolha dos substantivos com que se trabalhará é um dos passos cruciais para o desenvolvimento desta pesquisa, pois é desta forma como se consegue demarcar a área de trabalho. No início, pretendia-se trabalhar só como nomes que contassem com o semema [+dialógico], como *Debatte*, *Diskussion* ou *Gespräch*, mas a realização de testes preambulares com a metodologia pretendida revelou que os dados podiam ser muito homogêneos para os três nomes pelo facto de estarem próximos do ponto de vista semântico e de realização da estrutura actancial. Por este motivo e para atingir maior grau de variação no que diz respeito à análise de resultados, decidiu-se trabalhar com três substantivos que, embora façam parte do campo lexical da comunicação, tenham alguma diferença do ponto de vista do seu significado. Os três nomes selecionados para tal fim são *Bericht* (> relatório), *Diskussion* (> discussão) e *Frage* (> pergunta) e a seguir serão definidos brevemente tomando-se em consideração as suas proposições e o seu sentido:

- *Bericht*: deve-se salientar que o substantivo *Bericht* é utilizado amiúde para designar uma representação ou reprodução linguística factual e objetiva de um determinado evento ou de uma série de circunstâncias. Isto é, Hernández (1993, p. 131) assinala que com esta unidade lexical não se transmite informação de tipo subjetivo, senão que se trata de factos reais, detalhados ou inclusive oficiais. Trata-se, aliás, de um substantivo monológico dentro do campo lexical da comunicação, pois o emissor informa ao recetor sobre uma mensagem. Graças a esta estruturação do significado, pode-se afirmar que o agente (emissor) costuma contar com os traços [+humano] ou [+instituição]: *Der Bericht* ***des Rates*** (Sketch Engine. German Web 2018: fds-sprachforschung.de).
- *Diskussion*: normalmente, *Diskussion* corresponde a uma argumentação recíproca entre os participantes, na qual eles podem contribuir com as suas opiniões para um dado tema (Schumacher, 1986, p. 704). Aliás, neste caso, destaca-se a tendência para a controvérsia existente entre as pessoas envolvidas na discussão (Hernández, 1993, p. 94). Relativamente ao agente, ele pode estar representado por uma pessoa só ([+humano]) ou por um grupo (igualmente podendo ser [+institucional]): *Die Diskussion* ***der Wissenschaftler mit Politikern*** (Sketch Engine. German Web 2018: energy-harvesting-net.de)[35].
- *Frage*: no caso de *Frage*, trata-se de um substantivo que acostuma aparecer, ou ao menos assim está organizado no nosso léxico, em conjunção com *Antwort* (> resposta) (Hernández, 1993, p. 93). A pergunta age, portanto, como impulso que conduz ao surgimento de uma resposta. Não se podem descrever claramente como nomes dialógicos ou monológicos, pois costumam motivar a presença de, no mínimo, dois interlocutores no quadro geral do intercâmbio: *die Frage* ***des Moderators an den Reporter*** (Sketch Engine. German Web 2018: se-lp.de).

A seleção destes três substantivos pode-se justificar, aliás, pelo facto de serem lexemas em que o actante semântico do agente está muito presente nas estruturas argumentais, como será explicado com mais detalhe em § 6, sendo que até são possíveis diferentes realizações formais para a função de complemento sujeito. Nesta linha, Wierzbicka (1996, p. 178) salienta que a expressão explícita quer do locutor quer do alocutário acaba por ser necessária do ponto de vista sintático e do ponto de vista

[35] Tradução própria: "a discussão dos cientistas com os políticos".

semântico, especialmente no caso de lexemas dialógicos (como *Diskussion*), pelo facto de implicarem a presença simultânea e a participação ativa de duas ou mais pessoas ou agentes semânticos.

Assim, para os nomes apurados, em dicionários que se tomam frequentemente como ponto de partida para a investigação sobre estruturas argumentais nominais (cf. PORTLEX; Sommerfeldt & Schreiber, 1983b) observa-se agora que a ativação de *slots* semânticos e valenciais diferentes (complemento sujeito, complemento objeto...) pode ser realizada, no plano formal, com uma mesma estrutura. Por outras palavras, nota-se que estruturas formais e frequentes em alemão como as frases em genitivo são empregues, no nível formal, para a concretização de vários complementos nominais.

No caso de Sommerfeldt e Schreiber (1983b, pp. 141-142), se consultarmos o artigo lexicográfico para o nome *Diskussion*, observamos que um substantivo em genitivo, apareça ele em singular ou em plural, pode ser utilizado para a ativação de dois argumentos. Estes autores oferecem dois exemplos claramente demarcados: o primeiro diz respeito ao complemento objeto, pois recolhem o genitivo com o traço semântico [abstrato] fornecendo como ilustração a frase nominal *die Diskussion der Frage* com o sentido 'discute-se sobre um problema'; em segundo lugar, compilam a realização do genitivo com um substantivo [humano], sendo que o complemento agora realizado é o sujeito, exemplificando-se com a frase *die Diskussion des Jungen*, agora com o significado de 'o jovem discute'.

Por sua parte, no PORTLEX recolhem-se também dois tipos de complementos com a realização formal em genitivo para o lema *Diskussion.* Neste dicionário, estabelece-se que o genitivo pode ser empregue, em conjunção com o substantivo *Diskussion*, para ativar o actante 'aquele/aquilo que realiza uma ação' ou 'aquele/aquilo não afetado: tema'. Assim, no exemplo que se fornece para primeiro caso, *die Diskussion der Abgeordneten*, são os deputados que discutem, enquanto para o segundo caso, na frase *die Diskussion des Arbeitsprogramms*, verifica-se que alguém (não mencionado explicitamente na estrutura) discute sobre o programa de trabalho.

Daí deduz-se que a principal diferença que existe entre os dois complementos aqui abordados se encontra no nível semântico, tanto no nível do significado relacional como no nível do significado categorial. Destarte, tal como apontam Domínguez *et al.* (2018) ou López (2020), as consultas CQL baseadas em parâmetros sintáticos ou formais, denunciam-se como insuficientes para uma extração ausente de ruído pela inexistência de anotação semântica. De acordo com Jakubíček *et al.* (2010, p. 742), o formalismo que acarreta a etiquetagem sistemática seguindo parâmetros da língua de marcação XML possibilita uma procura bem-sucedida, mas isto acontece se se considerarem somente critérios formais e sintáticos. Por exemplo, a delimitação de uma consulta CQL como a seguinte,

`[lemma="Diskussion"] [tag="(ART\.(Def|Indef) | PRO.(Dem|Poss).Attr).Gen.*"] [tag="N.*"]`, extrai, quando é utilizada em Sketch Engine, dados que correspondem aos dois padrões valenciais, não possibilitando logo uma desambiguação ou distinção rápida entre as frases onde o genitivo é complemento sujeito e aquelas onde funciona como complemento objeto, pois aparecem estruturas como *Diskussion der Ergebnisse* (discute-se sobre os resultados) junto com outras como *Diskussion der Kandidaten* (os candidatos discutem)[36].

Desta forma, verifica-se que se pretendem extrair do *corpus* várias concordâncias em que um nome *n* funcione como predicado e partir-se-á, para tal fim, de uma determinada realização formal, como o genitivo. Considerando-se os aspetos *supra* referidos, podem-se retirar os dados anotados em linguagem de marcação XML a partir do Sketch Engine (*vd*. código *infra*), o que facilita a futura manipulação da informação através do *script* informático que se desenhará *ad hoc*. As consultas CQL aparecem anotadas de forma automática com a etiqueta `<kwic>` (*Key Word in Context*)[37], de modo a podermos aplicar métodos posteriores só àquela parte que foi marcada *a priori*:

```
<?xml version='1.0' encoding='UTF-8' ?>
<export>
<header>
<corpus>preloaded/detenten18_rft3</corpus>
<subcorpus>-</subcorpus>
<query>
<subquery operation="Query"
size="10227">[lemma=&amp;quot;Diskussion&amp;quot;][tag=&amp;quot;(ART\.(Def|Indef)|PRO
.(Dem|Poss).Attr).Gen.*&amp;quot;][tag=&amp;quot;N.*&amp;quot;]</subquery>
</query>
</header>
<concordance>
<line refs="ulrich-schrader.de" num="0" label="_">
<left>Bewertung oder</left>
<kwic>Diskussion der Ergebnisse</kwic>
<right>findet hier nur</right>
</line>
<line refs="josberlin.de" num="1" label="_">
<left>werden. &lt;/s&gt;&lt;s&gt; Eine</left>
<kwic>Diskussion der Notwendigkeit</kwic>
```

[36] Ao longo deste capítulo, mencionam-se os complementos sujeito e objeto de forma específica porque se estão a fornecer sempre exemplos relativos à estrutura argumental com a unidade lexical *Diskussion* sendo o predicado. No presente trabalho, partir-se-á, não obstante, da realização formal, isto é, de frases em genitivo, para a desambiguação de argumentos, seja qual for o tipo específico em que se enquadram. Mais informação sobre a realização destes actantes será providenciada para cada nome em § 6.

[37] Do ponto de vista de PLN, afirma-se que se trata, neste caso, da etiqueta XML para as concordâncias de um nome em Sketch Engine.

```
<right>und Zulässigkeit</right>
</line>
<line refs="weltladen-freising.de" num="2" label="_">
<left>gesammelt haben &lt;/s&gt;&lt;s&gt;</left>
<kwic>Diskussion der Kandidaten</kwic>
<right>&lt;/s&gt;&lt;s&gt; Diskussion des</right>
</line>
```

Para a análise posterior, serão efetuadas provas preliminares com dados de frequência absoluta em *corpora* para cada substantivo, pois parte-se da hipótese de que, tal como afirmam Domínguez *et al*. (2021), os critérios de frequência nem sempre podem ser considerados como essenciais quando são analisados padrões valenciais, pois algumas realizações formais de actantes semânticos ou sintáticos aparecem na língua natural, mas não ocupam uma posição relevante em *corpora* organizado seguindo exclusivamente parâmetros estatísticos de frequência.

Para além disso, devido ao facto de o foco desta pesquisa se centrar na etiquetagem semântica, criar-se-ão pacotes lexicais *ad hoc* que permitam uma anotação automática com os traços semânticos selecionados, por exemplo, com o traço [humano]. Tomando esta primeira abordagem como ponto de partida e protótipo, poder-se-ia desenhar uma metodologia sistematizada de anotação automática em *corpora*. Como se explicou em § 2.3. (*vd*. *supra*), utiliza-se a ontologia desenhada no âmbito dos projetos MultiGenera e MultiComb como alicerce, embora precise de ser adaptada para os objetivos aqui perseguidos.

Neste sentido, para a criação do pacote lexical que subjazerá ao *script* empregue no presente trabalho e que funcionará como *input* para o sistema de anotação, alargar-se-á o conceito de pacote lexical definido por Domínguez *et al*. (2021). Consoante Domínguez *et al*. (2021, p. 277), um pacote lexical está formado por unidades lexicais relacionadas que, embora não sejam sempre intercambiáveis entre si, têm uma relação próxima no eixo paradigmático. De modo a ilustrar esta explicação, afirma-se que unidades lexicais como *Bruder*, *Ehepartner*, *Enkelin* ou *Halbschwester* fazem parte do pacote lexical restrito {animado, humano, família}.

No começo deste projeto, pretendia-se atingir uma anotação semântica específica para cada pacote lexical concreto[38], mas a obtenção de um número muito reduzido de frases anotadas alterou a aproximação ao *corpus*. Destarte, um teste preliminar com o pacote lexical {animado, humano, profissão, educação} onde só foram anotadas semanticamente de forma automática aproximadamente cinco frases

[38] Isto implica que se queria começar com a automatização da anotação semântica partindo das últimas categorias da ontologia: {animado, humano, profissão}, {animado, humano, família}, {animado, humano, origem}, etc.

(isto equivale a uma percentagem de 0,06% do total), fez com que se decidisse ampliar o pacote para {animado, humano}. A designação que se empregará na presente pesquisa para nomear o pacote lexical utilizado, ou seja, {animado, humano}, é mais abrangente que a noção de [humano] empregada em PORTLEX ou em Sommerfeldt e Schreiber (1983b), pois sob a nossa classificação incluem-se também organizações regidas por humanos, que, do ponto de vista dos autores anteriores, correspondem à etiqueta de [instituição]. Deste modo, confirma-se que vocabulário que se define como {animado, humano, organização de ensino} aparece, por exemplo, como fazendo parte do nosso pacote lexical {animado, humano}. Consoante o defendido por Renau e Nazar (2016a, p. 824), nem sempre é necessário contarmos com uma taxonomia lexical muito detalhada, sendo que por vezes se prefere operar com classes semânticas mais amplas para identificar determinados usos linguísticos desprovidos de elevado detalhe lexical.

Para além de se juntarem unidades lexicais de diferentes pacotes prototípicos provenientes dos projetos MultiComb e MultiGenera, realizou-se um processo de pré-processamento de modo a evitar repetições desnecessárias de palavras e de modo a conseguir uma lista de vocabulário arrumada e desprovido de ruído. Para tal fim, lançou-se mão, entre outras ferramentas e editores de texto, do comando `sort -u` no interpretador da linha de comando de Unix. Como consequência deste procedimento, obteve-se uma lista de vocabulário que se utilizará em fases ulteriores como *input* para o *script*. Destaca-se que nesta listagem, os substantivos não aparecem lematizados, senão que se utilizam diferentes realizações morfossintáticas das unidades lexicais registadas. Neste sentido, podemos observar que costumam ocorrer no pacote lexical as formas de genitivo singular e plural, para além de algum nominativo ou dativo[39], pois com o recurso aos lemas tornar-se-ia necessário utilizar de forma constante expressões regulares que possibilitassem o processo de anotação pretendido. A utilização de palavras morfossintáticas em vez de lexemas permite uma anotação mais exaustiva, pois o *script* só compara o pacote lexical pré-desenhado com o ficheiro XML inserido como *input* para a etiquetagem semântica. Aliás, note-se que não se introduziram nomes próprios dentro da taxonomia de trabalho pela grande variedade deles que existe e a dificuldade para os sistematizar (cf. Renau & Nazar, 2016a).

**Tabela 2**

*Lista de nomes extraídos do pacote lexical {animado, humano}*

| | | |
|---|---|---|
| Abendländer | Abenteurerinnen | Aborigine |
| Abendländern | Abenteurern | Aborigines |

[39] Embora a lista seja destinada à anotação de genitivos fundamentalmente neste trabalho, o pacote lexical inclui outras realizações morfossintáticas de modo a não restringir a sua aplicabilidade.

| | | |
|---|---|---|
| Abendländers | Abenteurers | Absender |
| Abenteuerin | Abiturient | Absenderin |
| Abenteuerinnen | Abiturienten | Absenderinnen |
| Abenteurer | Abiturientin | Absendern |
| Abenteurerin | Abiturientinnen | Absenders |

*Nota*. Este pacote lexical de trabalho foi desenhado no âmbito deste projeto, mas deve-se considerar que se tomaram os pacotes preexistentes dos projetos MultiComb e MultiGenera. Agradeço imenso à Prof.ª Dr.ª María José Domínguez, diretora dos projetos antes citados, pela disponibilização dos pacotes em língua alemã para o desenvolvimento desta investigação.

Tendo-se constituído o pacote lexical {animado, humano} com o qual se trabalhará na presente dissertação, torna-se necessário desenvolver um *script* informático *ad hoc* para atingir a anotação pretendida das unidades lexicais. Decidiu-se operar com a linguagem de programação Perl pelo facto de ser apropriada para a manipulação de textos, permitindo realizar substituições que podem facilitar a anotação perseguida. Aliás, na primeira linha do *script* deve-se incluir a declaração de execução, para a interface de Unix reconhecer o programa escrito em linguagem Perl. A seguir, apresentar-se-á o script desenvolvido de forma mais pormenorizada, fornecendo-se informações sobre o processo de criação e de execução[40]:

```
#!/usr/bin/perl
use strict;
my @hum=();
while(<DATA>){
   chomp;
   push(@hum,$_);
}
my $pattern = join("|",@hum);
while(<>){
   next unless /kwic/;
   s!\b($pattern)\b!<sem_tag type="human">$1</sem_tag>!g;
   print;
}
__DATA__
Abendländer
Abendländern
Abendländers
Abenteuerin
Abenteuerinnen
Abenteurer
```

[40] Neste ponto, quero agradecer ao Prof. Dr. José João Almeida pela ajuda fornecida para o desenvolvimento do *script* aqui apresentado.

Com o *script* apresentado acima, pretende-se atingir uma anotação XML[41] dos documentos fornecidos como *input* no sistema, de modo que o *output* alcançado tenha uma etiquetagem semântica das partes de interesse. Para tal fim, deve-se declarar no próprio programa que só se utilizará em modo estrito (`use strict`), isto é, com as variáveis aí estabelecidas. Neste caso, a variável de tipo *array* (`my @hum`) é formada por todos os valores introduzidos na parte inferior, sob a linha `__DATA__` indicando-se que toda a informação a seguir é fornecida como *input* ao programa. A função da variável `@hum` acaba por consistir em juntar (`|`) todas as unidades lexicais inseridas em `__DATA__` para posteriormente verificar se elas aparecem nos ficheiros XML. Isso é realizado várias vezes, enquanto a condição for certa, graças ao *loop* `while`. Para além disto, pretende-se ficar só com as linhas anotadas previamente com o *tag* `<kwic>` para ver se ocorrem nelas algumas das palavras que tem de ser etiquetadas semanticamente. Para tal propósito, as ocorrências das unidades lexicais elencadas sob `__DATA__` serão encontradas e substituídas por elas mesmas acompanhadas de um *tag* em linguagem de marcação XML formado pelo elemento `<sem_tag>` e pelo atributo `type`. Tenciona-se, aliás, que o `type` seja sempre `"human"` pelo facto de estarmos a trabalhar com o pacote lexical {animado, humano}.

Após o *script* ser desenvolvido, ele deve ser executado na linha de comandos de Unix como um programa Perl, para o que se deve utilizar o comando `perl` seguido do nome do programa, neste caso, `annotate.pl`. Assim, também se deve introduzir o nome do documento que funcionará como *input*, mais concretamente, o ficheiro XML descarregado do Sketch Engine para um dado nome e fazer com que o *output* seja redirigido (`>`) para outro documento, de modo a termos guardado o ficheiro anotado. Evidencia-se, aliás, que as consultas CQL, embora só funcionem no nível formal e morfossintático, facilitam enormemente a manipulação posterior, uma vez que podemos depurar previamente o ficheiro e anotar somente as palavras que constituem o alvo da nossa pesquisa.

Antes de começar com a análise dos três substantivos apurados, assim como das suas estruturas argumentais para atingir uma anotação e posterior extração semântica, torna-se crucial pormos à prova o programa Perl desenvolvido. Para tal objetivo, trabalha-se com as frases geradas automaticamente com a ferramenta Xera apresentada em § 4, sendo que só se escolhem estruturas mono-argumentais em que o actante presente diz respeito ao agente ou ao complemento sujeito, 'aquele/aquilo que realiza a ação', para alcançar a anotação semântica com a classe {animado, humano}. Alguns dos resultados obtidos apresentam-se a seguir:

[41] Para a validação externa do ficheiro XML, torna-se necessário criar um documento DTD (*Document type definition*) que inclua declarações de marcação. Esta parte técnica não será tematizada na presente dissertação.

<kwic>eine Diskussion der <sem_tag type="human">Abteilungsleiterinnen</sem_tag></kwic>

<kwic>diese Diskussionen der <sem_tag type="human">Abteilungsleiterinnen</sem_tag></kwic>

<kwic>diese Diskussion der <sem_tag type="human">Amtsinhaber</sem_tag></kwic>

<kwic>keine Diskussionen der <sem_tag type="human">Amtsinhaber</sem_tag></kwic>

<kwic>diese Diskussion der <sem_tag type="human">Amtsinhaber</sem_tag></kwic>

<kwic>die Diskussionen der <sem_tag type="human">Amtsinhaber</sem_tag></kwic>

<kwic>eine Diskussion der <sem_tag type="human">Amtsinhaberinnen</sem_tag></kwic>

<kwic>keine Diskussionen der <sem_tag type="human">Amtsinhaberinnen</sem_tag></kwic>

Corrobora-se, portanto, que o programa desenhado *ad hoc* está a funcionar de acordo com os propósitos perseguidos, pois como se pode ver nas concordâncias *supra*, localiza as unidades lexicais correspondentes à classe semântica {animado, humano} e anota-as automaticamente com o *tag* desenhado. Em consequência, poderá ser aplicado com maiores quantidades de texto, sendo que as estruturas argumentais devem estar anotadas previamente com um *tag* `<kwic>` para facilitar a sua manipulação. De modo a responder as questões de investigação apresentadas na secção 5.1., em § 6 trabalhar-se-á com o programa para anotar os ficheiros subjacentes à presente dissertação para cada um dos substantivos selecionados. Assim, visa-se descobrir a percentagem de exatidão do *script* e tenciona-se verificar se compensaria uma sistematização da anotação semântica com o pacote lexical {animado, humano} para atingir maior facilidade de desambiguação ou de extração de informação linguística.

# 6

## Apresentação e análise dos resultados

No presente capítulo apresentam-se os resultados obtidos nesta dissertação e com tal intuito, realizar-se-á uma avaliação do *script* não só em termos qualitativos, mas também em termos quantitativos através do recurso a aplicações estatísticas. Destarte, pretende-se fornecer, por um lado, informação de carácter linguístico e lexicográfico que possa servir para atingirmos uma descrição mais fidedigna dos três substantivos com que se trabalha. Por outro lado, deve-se recordar a importância de dar uma resposta baseada em dados empíricos às perguntas de investigação *supra* formuladas (*vd.* subcapítulo 5.1.), nas quais se destaca a importância de estimarmos a valia que supõe o desenvolvimento de um *script* informático para a anotação semântica (*vd.* subcapítulo 5.2.). Para tal finalidade, serão analisados os dados relativos a cada um dos substantivos apurados, pois é desta forma como se conseguirão observar tendências qualitativas e quantitativas e, em última instância, avaliar o funcionamento e viabilidade do *script*.

### 6.1. O substantivo *Bericht*

No que diz respeito ao substantivo alemão *Bericht*, há que salientar que todos os dicionários consultados para a presente pesquisa, afirmam que se trata de uma unidade lexical com uma só aceção. Esta consistência nos diferentes recursos lexicográficos examinados observa-se a seguir:

„(offizielle) Wiedergabe eines Geschehens, Sachverhalts“ (DWDS, n. d.)

„sachliche Wiedergabe eines Geschehens oder Sachverhalts; Mitteilung, Darstellung“ (DUDEN, n. d.)

„Wiedergabe eines Sachverhalts“ (Sommerfeldt & Schreiber, 1983b, p. 121)

Embora não se possa falar em monossemia *stricto sensu*, pelo facto de o substantivo poder adquirir diferentes sentidos ou, quando menos, nuances de significado ao ser combinado com outras unidades lexicais, afirma-se que a variação na estrutura argumental só se revela dependente de o significado *supra* representado. Assim, a definição lexicográfica ou a paráfrase do significado é construída, neste caso, através do recurso ao hiperónimo *Wiedergabe*, centrando-se o foco ainda nalguns dos aspetos abordados na descrição introduzida em § 5.2. da presente pesquisa. Ressalta-se, desta

forma, que se trata de uma unidade lexical cujo conteúdo semântico diz respeito à transmissão de uma mensagem meramente objetiva e factual, tal como defendido por Hernández (1993).

Para além da definição, deve-se apontar para a importância dos exemplos lexicográficos, pois do ponto de vista da metalexicografia, "eles podem ser muito ricos em informação gramatical, enciclopédica, pragmática ou sobre combinatória lexical" (Iriarte, 2001, p. 328). Para os objetivos desta dissertação, parece que os exemplos relativos à combinatória lexical da unidade lexicográfica são relevantes, pois pode-se entender sob a noção de combinatória qualquer combinação de unidades lexicais no plano sintagmático. Neste caso, segundo Iriarte (2001, p. 26), tratar-se-ia de combinatória lexical livre pelo facto de "a combinação das unidades lexicais [ser] feita segundo as regras gramaticais de uma língua".

Em consequência, se consultarmos os exemplos introduzidos nos artigos lexicográficos para o verbete *Bericht* nos diferentes dicionários sob consulta, observamos que alguns ajudam para a codificação linguística (Iriarte, 2004), sendo que fornecem informação sobre a estrutura argumental do substantivo lematizado, mesmo que não seja explicitado do ponto de vista gramatical. Destarte, a partir de exemplos como *der zusammenfassende Bericht des Sachverständigen* ou *die amtlichen Berichte der Regierung* (DWDS) pode-se deduzir que a estrutura em genitivo (*des Sachverständigen* ou *der Regierung*) é utilizada em alemão para esclarecer quem é 'aquele que realiza a ação', isto é, aqui, o complemento sujeito (cf. Domínguez, 2014). Igualmente, o dicionário DUDEN (n. d.) inclui informações a respeito das preposições ou, em sentido mais amplo, das frases prepositivas, que acompanham o nome *Bericht*, sendo que se especifica que este substantivo pode reger as preposições *von*, *über* e *zu*, embora não se desenvolva a tipologia de entidades ontológicas que podem ocorrer com estas estruturas. Para além disso, deve-se destacar que estes dicionários não costumam fornecer informação explícita sobre os casos gramaticais que acompanham as preposições[42], de modo que os dados introduzidos no artigo lexicográfico podem ser insuficientes para alguns utentes, como, por exemplo, os aprendentes de língua estrangeira.

A partir da observação dos exemplos lexicográficos já referidos, assim como da consulta de dicionários baseados exclusivamente em parâmetros valenciais (PORTLEX; Sommerfeldt & Schreiber, 1983b), oferecer-se-á uma estrutura argumental do nome *Bericht* para encetarmos a análise pretendida com fundamento linguístico. Para tal fim, utilizar-se-ão as descrições semânticas dos actantes empregues no dicionário PORTLEX e a classificação de argumentos proposta por Domínguez (2014b). Cabe lembrar

[42] Embora esta informação esteja implícita na formulação de alguns exemplos lexicográficos (cf. DUDEN, n. d., *Berichte über das Tagesgeschehen*), uma clarificação do ponto de vista gramatical (*über* + acusativo) contribuiria para a consolidação destes dados por parte de utentes [estrangeiros] cuja língua primeira não seja o alemão.

que na estrutura argumental só os argumentos (*Ergänzungen*) acabam por ser representados, sendo que os circunstantes (*Angaben*) podem apresentar-se quer do ponto de vista sintático quer do ponto de vista semântico com multiplicidade de realizações e que, aliás, eles não são específicos das estruturas argumentais de um dado substantivo, verbo ou adjetivo. No caso do nome *Bericht*, costuma aparecer no esquema argumental a seguinte lista de argumentos:

**Tabela 3**

*Estrutura argumental do nome* Bericht

| | |
|---|---|
| **Complemento sujeito**<br>'aquele/aquilo que realiza a ação'<br>[humano] [instituição] | Genitivo: *Der Bericht* ***des Vorstandes*** (Sketch Engine. German Web 2018: aekwl.de)<br>Frase preposicional *von*: *Berichte* ***von Mitgliedstaaten*** (Sketch Engine. German Web 2018: datenschmutz.de)<br>Determinante possessivo: ***Deine*** *Berichte* (Sketch Engine. German Web 2018: dampfradioforum.de)<br>Palavra composta: ***Augenzeugen****berichte* (Sketch Engine. German Web 2018: woz.ch)<br>Adjetivo: ***tschechische*** *Berichte* (Sketch Engine. German Web 2018: schoenhengstgau.de) |
| **Complemento objeto**<br>'aquele/aquilo não afetado: tema'<br>[imaterial] | Genitivo: *Bericht* ***der Reise*** (Sketch Engine. German Web 2018: fvms.de)<br>Palavra composta: ***Wetter****bericht* (Sketch Engine. German Web 2018: sy-arion.de)<br>Adjetivo: *ein* ***wirtschaftlicher*** *Bericht* (Sketch Engine. German Web 2018: kjr-myr.de) |
| **Complemento prepositivo**<br>'aquele/aquilo não afetado: tema'<br>[imaterial] | Frase preposicional *über*: *Bericht* ***über diese Veranstaltung*** (Sketch Engine. German Web 2018: garagensound.de)<br>Frase preposicional *von*: *Bericht* ***von der Jahrestagung*** (Sketch Engine. German Web 2018: nakos.de)<br>Frage preposicional *zu*: *Bericht* ***zu den ukrainischen Wahlen*** (Sketch Engine. German Web 2018: andigross.ch) |

| | |
|---|---|
| **Complemento prepositivo**<br>'aquele/aquilo não afetado: tema'<br>[situação] | *darüber* + oração subordinada completiva: [...] *Bericht* ***darüber, wie seine Arzneimittel entstehen und in die Hand des Arztes gelangen*** [...] (Sketch Engine. German Web 2018: hoimar-von-ditfurth.de) |
| **Complemento prepositivo**<br>'aquele/aquilo não afetado'<br>[humano] [instituição] | Frase preposicional *an*: *Berichte* ***an den Gutachter*** (Sketch Engine. German Web 2018: dcomnet.com) |

*Nota.* A estrutura argumental proposta para o nome *Bericht* foi realizada através do recurso fundamental ao dicionário PORTLEX e ao dicionário de Sommerfeldt e Schreiber (1983b), para além da consulta de *corpora* em Sketch Engine.

A partir da análise *supra* realizada em que se consideraram as diferentes realizações formais da estrutura argumental do nome *Bericht*, deduz-se que algumas delas podem revelar-se como ambíguas[43], pelo facto de uma mesma realização superficial servir para a ativação de, por exemplo, dois argumentos diferentes. Em seguida, serão apresentadas todas as estruturas linguísticas que são ambíguas no esquema actancial do substantivo alemão *Bericht*, prestando-se especial atenção aos aspetos relevantes para a aplicação ulterior do *script* informático explicado no subcapítulo 5.2.:

- Genitivo: a realização formal através do recurso a frases nominais em caso genitivo destaca-se como sendo uma das estruturas linguísticas mais frequentes para a realização do actante 'aquele/aquilo que realiza a ação', isto é, para o traço semântico do agente, embora também possa aparecer como realização superficial para o complemento objeto. A execução de uma consulta CQL em Sketch Engine como a seguinte, `[lemma="Bericht"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"][tag="N.*"]`, não permite clarificar o tipo de complemento de que se trata, uma vez que ambos os actantes apresentam a mesma constituição sintática, sendo que o único traço distintivo se encontra no nível semântico. Assim, alguns exemplos que obtemos como resultado da consulta CQL evidenciam este comportamento:
    - *Das geht aus dem* ***Bericht der Bundesregierung*** *hervor*. (Sketch Engine. German Web 2018: brh-nrw.de)

[43] Neste sentido, e de acordo com Móia (2016, pp. 309-310), pode-se falar nestes casos em "ambiguidade sintática atinente à estrutura em constituintes". Esclarece-se, portanto, que não se trata da ambiguidade semântica mais tradicional, senão que se considera como ambiguidade sintática onde uma estrutura linguística encaixada numa realização formal maior pode apresentar, no mínimo, dois significados diferentes. É desta forma como se deve entender a ambiguidade na presente dissertação.

- *Es wird ein **Bericht unserer Verhandlungssituation** geben und den aktuellen Stand der Projektentwicklung*. (Sketch Engine. German Web 2018: brh-nrw.de)
- *Das wird im **Bericht des Finanzdepartments** deutlich*. (Sketch Engine. German Web 2018: tagesanzeiger.ch)

- Frase preposicional *von*: de acordo com Engel (2004, p. 295), a realização formal através do genitivo para o complemento sujeito pode ser amiúde substituída por uma frase preposicional com a preposição *von*. Embora o dicionário de Sommerfeldt e Schreiber (1983b) só contemple para a preposição *von* a possibilidade de ativação do traço semântico 'aquele/aquilo não afetado: tema', uma consulta CQL como esta `[lemma="Bericht"][lemma="von"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Dat.*"]?[tag="N.*"]` em Sketch Engine permite-nos deduzir que existem mais possibilidades, nomeadamente a realização do complemento sujeito em frases como *Bericht von Migrantinnen* ou *Bericht von Eltern*. Isto pode exemplificar-se com a inclusão de alguns dos resultados obtidos com a consulta CQL:
  - ***Einen Bericht von einer Schneeschuhtour** kann man hier nachlesen*. (Sketch Engine. German Web 2018: weinviertel.net)
  - *Nach **Berichten von Ärzten** seien die häufigsten Verletzungen, die versorgt werden müssen, großflächige Brandwunden*. (Sketch Engine. German Web 2018: arche-nova.org)
  - *Wir stellen dabei auch **Berichte von Leuten** vor, die die Befreiung von Auschwitz miterlebt haben*. (Sketch Engine. German Web 2018: buendnis-toleranz.de)
- Composto: no caso dos lexemas compostos em alemão, importa salientar que alguns já não podem ser entendidos como parte da estrutura argumental, senão que já contam com um maior grau de lexicalização na língua. Aliás, a consulta CQL realizada, `[lemma=".*bericht"]`, encontra todas as palavras compostas cujo último lexema é a unidade lexical *Bericht*. Nesta linha, sistemas de informação lexical, como o OWID do Instituto da Língua Alemã, já recolhem algumas destas unidades lexicais como lemas. Tal é o caso de *Erfahrungsbericht* ou *Wetterbericht*, pois embora estas palavras possam ser

ainda analisadas graças à estrutura argumental[44], elas já estão fixadas na língua alemã. Assim sendo, a consulta CQL providencia-nos exemplos como os seguintes:

- *Wer gerne **Reiseberichte** liest, sollte die von Sven unbedingt lesen- Es ist fast als wäre man selbst dabei*. (Sketch Engine. German Web 2018: fuldaforum.de)
- *Laut einem **Zeitungsbericht** ist nun auch eine „Zelda"-App geplant*. (Sketch Engine. German Web 2018: insuedthueringen.de)
- ***Augenzeugenberichte** und Videos lassen den Schluss zu, dass sie mit Knüppeln und Pistolen vorgingen*. (Sketch Engine. German Web 2018: woz.ch)

- Adjetivo: Domínguez (2011, p. 142) destaca o facto de os adjetivos poderem ocupar uma posição actancial na frase nominal, uma vez que a autora afirma, utilizando como exemplo a frase *a dor de estômago*, que o complemento *de estômago* pode ser trocado pelo adjetivo relativo *estomacal*, cumprindo-se, desta forma, a mesma função sintática e semântica. No caso de *Bericht*, uma consulta CQL como `[tag="ADJA.*"][lemma="Bericht"]` permitiu-nos extrair todos os adjetivos em posição atributiva do Sketch Engine, sendo que se revelou tarefa complicada apurar adjetivos a preencherem um *slot* valencial, pois é frequente os adjetivos agirem como modificadores (cf. Arias, 2020) ou como colocativos (cf. Mel'čuk, 2015). Os resultados da consulta CQL evidenciam estas tendências:
  - *Ein **ausführlicher Bericht** folgt in unserer morgigen Ausgabe*. (Sketch Engine. German Web 2018: phorumursellis.de)
  - *Dies geht aus einem **offiziellen Bericht** der Polizei hervor*. (Sketch Engine. German Web 2018: 1blu.de)
  - *Ich finde es immer wieder sehr interessant, wie Leute **unterschiedliche Berichte** über den identischen Lauf präsentieren*. (Sketch Engine. German Web 2018: 100mc.de)

Deduz-se, em consequência, que todas estas estruturas linguísticas podem destacar-se como sendo as mais ambíguas na estrutura argumental do nome *Bericht* pelo facto de estarem registadas como realizações formais possíveis para mais de um actante. Deve-se lembrar, aliás, que embora não seja pertinente destacar a realização superficial de adjuntos ou circunstantes no âmbito da estrutura

[44] Deve-se destacar que, do ponto de vista da estrutura argumental, *Erfahrungsbericht* é equivalente à estrutura frasal *Bericht über Erfahrungen* ('relatório sobre experiências') e *Wetterbericht* corresponde com *Bericht über das Wetter* ('relatório sobre o tempo').

argumental, eles podem utilizar os mecanismos linguísticos *supra* elencados no nível cotextual, sendo que a necessidade de desambiguação se torna ainda maior.

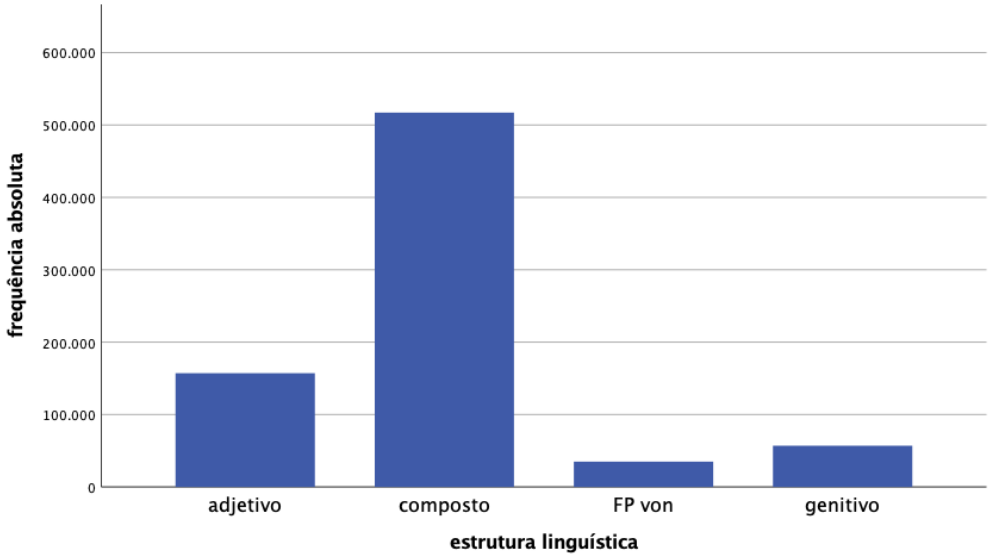
A figura 8 mostra a frequência absoluta no *corpus* de trabalho em Sketch Engine para as estruturas linguísticas que podem aparecer no padrão argumental de *Bericht* como realizações formais para diferentes argumentos. Aprecia-se, portanto, que os lexemas compostos constituem a estrutura actancial mais recorrente, visto que alguns estão muito lexicalizados na língua alemã como se referiu *supra*. Assim, as frases adjetivas também são mais habituais que outras estruturas devido ao seu carácter mais abrangente, pois como assinalado em Arias (2020), os adjetivos em posição atributiva[45] podem realizar muitas funções para além de serem ativadores de argumentos. De facto, Arias (2020) aponta para o facto de se tratar fundamentalmente de adjetivos atributivos incluídos na classe INHALT ('conteúdo'), entre os que sobressaem *ausführlich* ou *detailliert*, os mais frequentes em combinação com o nome *Bericht*.

Se se tomarem em consideração somente as duas classes restantes, observa-se que a realização superficial do genitivo é, em termos de frequência absoluta, mais habitual que a frase preposicional com *von*. Para além do critério quantitativo, deve-se lembrar que autores como Sommerfeldt e Schreiber (1983b) não incluem ainda a realização da frase preposicional *von* como parte da estrutura argumental para o complemento sujeito, embora os dados linguísticos agora extraídos e analisados a partir de *corpora* pareçam apontar para uma tendência clara segundo a qual esta realização formal é usada para a expressão do actante em questão.

[45] Consoante o exposto por Engel (2004), um exemplo de adjetivo atributivo é *die neue Ärztin*, onde o adjetivo aparece flexionado e ocorre antes do substantivo ao que se refere.

**Figura 8**

*Frequência absoluta em Sketch Engine: German Web 2018 das estruturas selecionadas*




Os critérios de frequência *supra* referidos, assim como a constituição do pacote lexical {animado, humano}, aspeto clarificado no subcapítulo 5.2., levaram para a seleção da estrutura de genitivo como a melhor candidata para a aplicação do *script* informático criado no âmbito da presente dissertação[46]. Destarte, visa-se oferecer respostas às perguntas de investigação subjacentes a esta pesquisa, assim como avaliar o funcionamento do protótipo de programa desenhado *ad hoc*. Para tal fim, parece razoável comentar a execução do *script* para uma determinada estrutura linguística, sendo que a sua utilidade deveria corroborar-se se aplicado noutras situações em que aparecesse o pacote lexical selecionado.

Todavia, parece comedido fazermos um pequeno excurso a critérios quantitativos de frequência antes de encetarmos a avaliação de carácter informático. Assim sendo, os dados relativos à frequência retirados do *corpus* de trabalho em Sketch Engine permitem-nos concluir que a unidade lexical *Bericht* costuma aparecer, na estrutura em genitivo, em combinação com substantivos que preenchem o *slot* do complemento sujeito e que podem, em consequência, ser anotados com o traço {animado, humano}. Não obstante, deve-se destacar também que a maior parte dos nomes listados na tabela 4 em conjunção com o lema *Bericht* contam com o traço [instituição], ou quando menos, [organização]: *Vorstand*, *Ausschuss*, *Bundesregierung*, *Komission*...

[46] Não obstante, esta abordagem não impede que o mesmo pacote lexical possa ser aplicado em pesquisas ulteriores com outras estruturas (como, por exemplo, von + {animado, humano}). Considera-se, neste trabalho, que o genitivo é a realização mais habitual para os traços semânticos humanos, pois embora os compostos tenham mais frequência em termos absolutos, uma análise qualitativa permite-nos logo deduzir que se trata em muitas ocasiões doutras realizações actanciais.

**Tabela 4**

*Dados mais frequentes para a estrutura* Bericht *+ genitivo em Sketch Engine: German Web 2018*

| lema | frequência |
|---|---|
| Bericht die Vorstand | 1615 |
| Bericht die Ausschuss | 1132 |
| Bericht die Kassenprüfer | 1104 |
| Bericht die Bundesregierung | 955 |
| Bericht die Kommission | 884 |
| Bericht die Vorsitzende | 635 |
| Bericht die Zeitung | 512 |
| Bericht die Aufsichtsrat | 396 |
| Bericht die Enquetekommission | 394 |
| Bericht die Bundesrat | 317 |
| Bericht die Nachrichtenagentur | 273 |
| Bericht die Tageszeitung | 245 |
| Bericht die Schatzmeister | 220 |
| Bericht die Kassenwart | 219 |
| Bericht die Arbeitsgruppe | 212 |
| Bericht die Nachrichtenmagazin | 207 |
| Bericht die New | 206 |
| Bericht die Landesregierung | 205 |
| Bericht die HNA | 204 |
| Bericht die Rechtsausschuss | 198 |
| Bericht die Magazin | 197 |

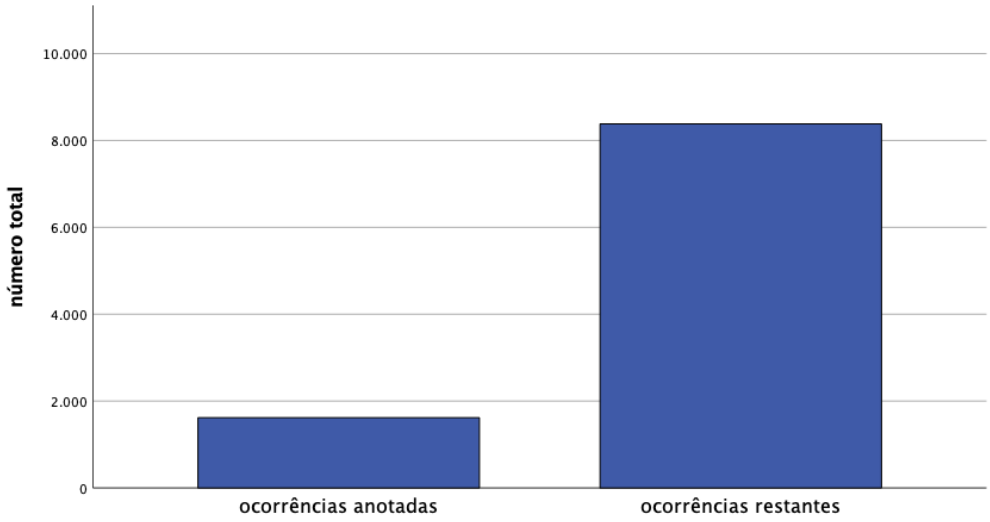
A partir do exposto na tabela 4, infere-se que a frase em genitivo é uma realização formal, pelo menos em termos de frequência, muito comum para o complemento sujeito em combinação com a unidade lexical *Bericht*, sendo que o 80% dos casos *supra* elencados podem ser classificados indubitavelmente com a etiqueta {animado, humano}. Não obstante, tal como assinalam Domínguez *et al*. (2021), a frequência não deve ser considerada como único critério para a análise de padrões valenciais. Neste sentido, conjetura-se que a anotação automática realizada graças ao *script* desenhado permitirá retirar conclusões mais decisivas a respeito das categorias semântico-ontológicss que acarreta a realização formal em genitivo.

A frequência absoluta, isto é, o número total de vezes que o *script* reconheceu e anotou uma unidade lexical como {animado, humano} é representada na figura 9 *infra*. Desta forma, pode-se apreciar a quantidade de ocorrências que foram etiquetadas de forma automática mediante a aplicação do *script*.

Deve-se clarificar que foram extraídas 10000 frases a partir do *corpus* utilizado em Sketch Engine[47], das quais 16,18% foram anotadas com o *tag* semântico com graças ao programa desenvolvido no âmbito desta dissertação. Apronta-se, em consequência, que o *script* consegue anotar as unidades lexicais no documento XML inserido como *input*, sendo que essas palavras devem fazer parte do pacote lexical com o qual se trabalha.

**Figura 9**

*Frequência absoluta das ocorrências anotadas como {animado, humano} para* Bericht




Esta anotação realizada através do *script* permite uma posterior extração, assim como uma fácil manipulação, dos dados linguísticos que foram previamente anotados com a etiqueta semântica. Assim, do ponto de vista da abordagem CPA apresentada no capítulo 4.2., pode-se afirmar que o padrão lexical que se extrai é o seguinte: *Bericht des/der* [[humano]]. De acordo com Hanks (2013), poderia delimitar-se ainda mais o padrão e estabelecer-se que o esquema estrutural extraído corresponde a *Bericht des/der* [[grupo humano]], sendo que se confirma a maior presença do traço semântico [instituição], tal como se indica em Sommerfeldt e Schreiber (1983b). Nesta linha, cabe destacar que as ocorrências mais frequentes entre as anotadas são as seguintes[48]:

```
222 <kwic>Bericht der <sem_tag
type="human">Kommission</sem_tag></kwic>

189 <kwic>Bericht der <sem_tag
type="human">Bundesregierung</sem_tag></kwic>
```

[47] A licença do Sketch Engine só permite descarregar um máximo de 10000 concordâncias para uma unidade lexical selecionada do *corpus* de trabalho.
[48] O número à esquerda de cada ocorrência corresponde à frequência em termos absolutos que aparece cada instância, sendo que só se considera a parte anotada semanticamente de forma automática.

```
167 <kwic>Bericht des <sem_tag
type="human">Ausschusses</sem_tag></kwic>

38 <kwic>Bericht des <sem_tag
type="human">Bürgermeisters</sem_tag></kwic>
```

Observando as ocorrências *supra* elencadas, deduz-se que o *script* funciona adequadamente e que consegue anotar automaticamente com um *tag* semântico aquelas unidades lexicais que aparecem no *corpus* e que estão simultaneamente recolhidas no pacote lexical {animado, humano}. Este passo representa um avanço importante para o tratamento e análise de estruturas argumentais, neste caso com o nome *Bericht*, pois para além da informação morfossintática, consegue-se peneirar através de um filtro puramente semântico.

Porém, também se puderam advertir alguns problemas provenientes do *script* que devem ser igualmente colocados na presente discussão. Trata-se, nomeadamente, das unidades lexicais que foram grafadas com hífen, sendo que em vários casos correspondem a nomes compostos em alemão. Nas ocasiões em que tanto a base do composto como a palavra determinante (*Basis* e *Bestimmungswort* na gramática alemã; cf. Engel, 2004) estão incluídas no pacote lexical, então são ambas as partes anotadas, originando uma situação de redundância, pois nestes casos bastaria com a base do composto (normalmente, o elemento situado à direita) ser anotada semanticamente, embora a palavra determinante contribua à constituição do significado do conjunto. O exemplo seguinte, retirado das entidades anotadas, permite-nos ilustrar esta falha do programa, sendo que a unidade lexical selecionada, *EU-Kommission*, foi duplamente anotada, pelo facto de estarem incluídas as duas partes separadas com hífen dentro do pacote lexical {animado, humano}[49]:

```
<kwic>Bericht der <sem_tag type="human">EU</sem_tag>-<sem_tag
type="human">Kommission</sem_tag></kwic>
```

Uma hipótese para resolvermos o problema apresentado seria incluir no pacote lexical unidades lexicais compostas e grafadas com hífen, embora este processo poderia acabar por ser interminável, devido ao facto de a língua alemã tender para a criação de novos compostos através da combinação de elementos. Todavia, deve-se chamar a atenção para o facto de este aspeto não causar muito ruído no resultado atingido, pois na maior parte das ocasiões só o segundo elemento do composto (a palavra base) aparece etiquetada, o que ajuda para a posterior extração de informação linguística e de padrões lexicais e valenciais:

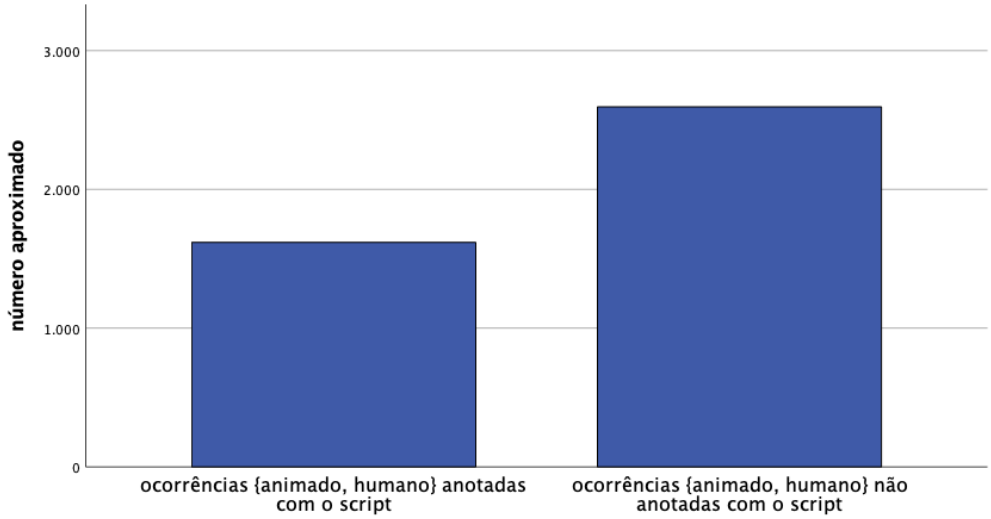
[49] Esta anotação dupla de que se fala não tem implicações significativas nos resultados finais obtidos, já que só uma percentagem reduzida de dados conta com duas etiquetas.

```
<kwic>Bericht der IAEA-<sem_tag type="human">Delegation</sem_tag></kwic>
```

Deve-se prestar atenção especial àqueles casos que não foram anotados automaticamente com a etiqueta semântica e que, por sua parte, podem ser classificados como {animado, humano} do ponto de vista lexical. Para esta parte da pesquisa, utilizou-se um foco semiautomático, sendo que o recurso a expressões regulares se tornou essencial para conseguir retirar conclusões definitivas acerca da não anotação de algumas entidades. Na gráfica da figura 10 *infra*, vê-se que uma percentagem aproximada de 62% das unidades que se correspondem com o padrão semântico {animado, humano} não foram anotadas com o *script* desenhado.

**Figura 10**

*Número aproximado de ocorrências {animado, humano} no corpus*




*Nota.* O número de ocorrências anotadas automaticamente com o *script* foi calculado também de forma automática. No entanto, as ocorrências não anotadas que podem ser classificadas como {animado, humano} foram contadas manualmente com a execução de expressões regulares.

A falta de anotação semântica de boa parte das ocorrências que podem ser catalogadas como {animado, humano} leva-nos a colocar a seguinte questão: qual é o motivo de esse conjunto de entidades não terem sido anotadas com o *script* automático? A utilização de expressões regulares, como já se referiu anteriormente, foi essencial para podermos indagar a ausência de etiquetagem em alguns casos. Em consequência, apresenta-se a seguinte relação de elementos não anotados automaticamente, assim como algumas das razões que podem ocasionar a falta de reconhecimento das entidades lexicais:

- Ausência de determinadas unidades lexicais no pacote lexical pré-desenhado: este é o principal motivo pelo qual algumas entidades não foram anotadas semanticamente de forma automática. O facto de algumas palavras não serem incluídas *a priori* no pacote

lexical de trabalho fez com que o *script* não as reconhecesse e anotasse em consequência. Nesta classe destacam-se, sobretudo, compostos em língua alemã para os quais somente a palavra base aparece no pacote lexical. Através do recurso a expressões regulares, conseguiram-se filtrar algumas destas ocorrências, pois a aplicação de expressões como `.*prüfer` ou `.*vorsitzende` permitiu-nos anotar mais informação linguística e semântica de forma manual.

- Presença de siglas preenchendo o *slot* semântico 'aquele/aquilo que realiza a ação': calculou-se que 9% das unidades que podem funcionar como agente no âmbito da estrutura argumental *Bericht des/der* [[humano]] correspondem com siglas. Isto tem a ver com o facto de o traço [instituição] estar muito presente na estrutura actancial deste substantivo (cf. Sommerfeldt & Schreiber, 1983b), assim como pelo facto de as siglas serem muito frequentes para nos referirmos aos nomes próprios de organizações ou empresas. Isto observa-se com exemplos como os seguintes: *Bericht der HNA* ou *Berichten des NDR*. Esta problemática evidencia, portanto, a necessidade de trabalhar com traços semânticos, pois corrobora-se que a retirada de dados a partir de *corpora* com critérios estatísticos pode conduzir a resultados inadequados quando se pretende elaborar recursos de orientação valencial.
- Presença de unidades lexicais difíceis de desambiguar: destacam-se, neste sentido, os casos de polissemia regular, pois segundo Apresjan (1974), trata-se de unidades lexicais com vários significados, embora eles estejam relacionados entre si através de relações de metonímia ou metáfora. Renau[50] (2021) analisa casos de polissemia regular em espanhol e assinala, aliás, que se revela, no âmbito linguístico, como um fenômeno habitual de criatividade lexical. Mais de 100 concordâncias extraídas do Sketch Engine contêm a frase *Bericht der Zeitung*, na qual o lexema *Zeitung* pode ter, quando menos, duas interpretações:
    - '[[instituição]] responsável pela publicação de um jornal'
    - '[[texto]] que resulta do processo de publicação periódica e contém notícias e informação de atualidade; jornal'

  Seguindo esta ideia, pode-se constatar que na frase *supra* citada cabem duas interpretações. A primeira diz respeito ao agente, pois pode-se entender que se trata de

[50] Cumpre destacar alguns trabalhos como o de Renau & Nazar (2016b), no qual se tenta desenvolver protótipos de algoritmos para procurar e analisar automaticamente os casos de polissemia regular.

um relatório criado pelo jornal, isto é, o jornal como sendo a instituição responsável pela realização da ação. Neste caso, as frases deveriam ser anotadas com o *tag* semântico. A segunda hipótese tem a ver com o jornal como 'texto' e, em consequência, como lugar físico onde um relatório é publicado.

- Ocorrência de nomes próprios: para alguns exemplos, entre os quais se podem mencionar algumas frases como *Bericht des Markus* ou *Bericht des Lukas*, cumpre salientar que nomes próprios, sejam eles antropónimos ou topónimos, podem acabar por preencher o *slot* semântico correspondente ao agente. Renau e Nazar (2016a) apontavam para a dificuldade de registar em ontologias e taxonomias nomes próprios, devido à grande quantidade deles que existe e a pouca uniformidade que apresentam dependendo da cultura de que se fala.
- Existência de erros tipográficos ou necessidade de pré-processamento: a presença de ruído devido à falta de uma fase estrita de pré-processamento das concordâncias faz com que algumas entidades não sejam anotadas por não corresponderem com a forma inserida no pacote lexical. Ocorrências como `<kwic>Bericht des Vorsitzenden3</kwic>` ou `<kwic>Berichte der Mannschaften3</kwic>` com a presença de números impede um reconhecimento claro e uma anotação semântica realizada pelo *script*.

Por sua parte, devem-se considerar outras situações que nos permitem avaliar o adequado funcionamento do *script* desenhado, entre as que cabe destacar as seguintes:

- Correspondência das entidades não anotadas com outro padrão lexical ou valencial: conforme descrito na estrutura argumental *supra* apresentada, o genitivo pode funcionar como realização formal para o complemento objeto. Neste caso, o padrão lexical corresponderia com a seguinte estrutura, *Bericht des/der* [[imaterial]], sendo que só se poderia atingir uma anotação semântica destas unidades lexicais criando um novo pacote lexical com lexemas que contassem com o traço semântico [imaterial]. Alguns dos exemplos encontrados listam-se a seguir: *Bericht der/einer/meiner Reise* ou *Bericht einer Entführung*.
- Correspondência das entidades não anotadas com realizações de adjuntos e circunstantes: o genitivo também pode ocorrer para desempenhar a função de um circunstante. Nas frases nominais *Bericht des Jahres* ou *Berichte dieser Woche* pode-se

apreciar que o genitivo possui a função de circunstante temporal, sendo que o substantivo em caso genitivo fornece informação relativa a um intervalo.

Em conclusão, as duas situações *supra* listadas, assim como o facto de termos unidades etiquetadas que efetivamente estavam presentes no pacote lexical aponta, em primeiro lugar, para um correto funcionamento do *script*. Não obstante, deve-se considerar que o pacote lexical era constituído por um número limitado de palavras, o que conduz à impossibilidade de anotar todos os lexemas de uma língua que contem com o traço semântico [humano]. Neste sentido, considera-se a possibilidade de retroalimentar o pacote lexical através da inserção de novas unidades lexicais que permitam tornar a ontologia de trabalho maior e, portanto, contribuir para o seu enriquecimento.

Por outro lado, deve-se lembrar que o facto de alguns elementos dos *supra* elencados não serem anotados aponta, igualmente, para o adequado funcionamento do programa desenhado, pois como já se viu, não são etiquetadas outras estruturas argumentais, isto é, outros complementos, de modo que se satisfazem alguns dos propósitos perseguidos, pois consegue-se uma desambiguação de estruturas argumentais. Assim, também se evita a anotação de estruturas com ruído do ponto de vista computacional. Conclui-se, em consequência, que o *script* está a funcionar de forma satisfatória para as estruturas frasais com o nome *Bericht*, mesmo que os dados linguísticos incluídos no pacote lexical possam continuar a ser aperfeiçoados.

### 6.2. O substantivo *Diskussion*

Relativamente ao substantivo *Diskussion*, nos dicionários monolingues em língua alemã que foram examinados para a presente dissertação, existe consenso na definição que reflete um 'intercâmbio de informação', embora o dicionário DUDEN (n. d.) acrescente outro significado mais restritivo, referido exclusivamente ao tratamento de temas diversos no âmbito da esfera pública (veja-se a segunda aceção):

„Meinungsaustausch, Auseinandersetzung" (DWDS, n. d.)

„1. [lebhaftes, wissenschaftliches] Gespräch über ein bestimmtes Thema, Problem

2. in der Öffentlichkeit (in der Presse, im Fernsehen, in der Bevölkerung o. Ä.) stattfindende Erörterung von bestimmten, die Allgemeinheit oder bestimmte Gruppen betreffenden Fragen" (DUDEN, n. d.)

„Meinungsaustausch" (Sommerfeldt & Schreiber, 1983b, p. 121)

Das paráfrases de significado *supra* elencadas deduz-se que todas apontam para um certo nível de harmonia nas diferentes fontes consultadas. Isto corrobora-se com o facto de ser o dicionário

DUDEN (n. d.) o único que contempla mais de uma aceção (polissemia) para o lema *Diskussion*, pois o DWDS só providencia uma definição, o que permitir-nos-ia falar em monossemia da unidade lexical. Para além deste tratamento lexicográfico, destaca-se que, tal como afirmava Schumacher (1986), o substantivo *Diskussion* corresponde, do ponto de vista do seu conteúdo semântico, a uma argumentação recíproca entre as pessoas envolvidas na conversa. Não obstante, note-se que o papel semântico do agente pode estar desempenhado através de unidades lexicais que ontologicamente tenham a ver com o traço [instituição] (*vd.* § 5.2.).

Se nos debruçarmos sobre os exemplos introduzidos no artigo lexicográfico quer do dicionário DUDEN (n. d.) quer da obra de consulta DWDS (n. d.), observa-se que é notória a ausência de informação fornecida acerca da construção dos padrões argumentais sintático-semânticos. De acordo com Zgusta (1971, p. 264), os exemplos lexicográficos devem servir para esclarecer a informação dada na definição, além de ilustrarem frequentemente o uso habitual de uma unidade lexical determinada.

Por um lado, o DUDEN (n. d.) oferece exemplos que incluem combinações lexicais do lema *Diskussion* com diferentes verbos suporte (cf. Mel'čuk, 2015; Iriarte, 2001), mas a inclusão de parâmetros valenciais formais e semânticos cinge-se à seguinte estrutura: *es gab, entbrannte eine leidenschaftliche, erregte Diskussion über, um den Paragrafen 218*. Desse esquema argumental, deduz-se somente que a unidade lexical *Diskussion* pode ser combinada com as preposições *über* e *um*, mas a ausência de informação relativa ao comportamento formal e semântico das preposições dificulta ainda a possibilidade de descodificar por parte de utilizadores de língua estrangeira. Informação muito similar é oferecida pelo DWDS, pois neste caso o exemplo *eine Diskussion über, um alle Probleme* só torna mais uma vez evidente que a combinação da unidade lexical com essas preposições é possível e correta gramaticalmente para transmitir o papel semântico de 'aquilo não afetado: tema', dados não especificados na obra lexicográfica.

Esta breve análise dos artigos lexicográficos do DUDEN (n. d.) e DWDS (n. d.) permite-nos concluir que a descrição dos padrões valenciais em dicionários de língua geral não é exaustiva nem suficiente para aprendentes de língua estrangeira, como já foi destacado por autores como Domínguez (2011). À volta da estrutura argumental da unidade lexical *Diskussion* há muita informação fornecida por dicionários de carácter valencial, como é o caso do PORTLEX ou de Sommerfeldt e Schreiber (1983). Baseando-nos nessas obras, assim como na consulta de *corpora*, realiza-se uma descrição mais pormenorizada dos padrões sintático-semânticos na tabela 5 a seguir. Para tal fim, utiliza-se (*vd.* § 6.1.) a nomenclatura de Domínguez (2014b) para a delimitação dos diferentes argumentos.

**Tabela 5**

*Estrutura argumental do nome* Diskussion

| | |
|---|---|
| **Complemento sujeito**<br>'aquele/aquilo que realiza a ação'<br>[humano] [instituição] | Genitivo: *Diskussion* ***der Politiker*** (Sketch Engine. German Web 2018: kremser.info)<br>Frase preposicional *unter*: *Diskussion* ***unter den Betreuern*** (Sketch Engine. German Web 2018: bfs-berlin.de)<br>Frase preposicional *von*: *die Diskussion* ***von Studenten aus Berlin*** (Sketch Engine. German Web 2018: tu-berlin.de)<br>Frase preposicional *zwischen*: *eine rege Diskussion* ***zwischen den Experten*** (Sketch Engine. German Web 2018: klagsverband.at)<br>Determinante possessivo: ***unsere*** *Diskussion* (Sketch Engine. German Web 2018: uebermedien.de)<br>Palavra composta: ***Gruppen****diskussion* (Sketch Engine. German Web 2018: kauzen.de)<br>Adjetivo: ***gesellschaftliche*** *Diskussion* (Sketch Engine. German Web 2018: zigarettenverband.de) |
| **Complemento objeto**<br>'aquele/aquilo não afetado: tema'<br>[imaterial] | Genitivo: *Diskussion* ***der Ergebnisse*** (Sketch Engine. German Web 2018: ulrich-schrader.de)<br>Palavra composta: ***Rechtschreib****diskussion* (Sketch Engine. German Web 2018: fds-sprachforschung.de)<br>Adjetivo: ***politische*** *Diskussionen* (Sketch Engine. German Web 2018: hanse.de) |
| **Complemento prepositivo**<br>'aquele/aquilo afetado'<br>[humano] [instituição] | Frase preposicional *mit*: *die Diskussion* ***mit der Ministerin*** (Sketch Engine. German Web 2018: uni-kiel.de) |
| **Complemento prepositivo**<br>'aquele/aquilo não afetado: tema' | Frase preposicional *über*: *eine Diskussion* ***über die Risiken der Bundesbank*** (Sketch Engine. German Web 2018: t-online.de) |

| | |
|---|---|
| [imaterial] | Frase preposional *um*: *die Diskussion* ***um die Verkehrsinfrastrukturen*** (Sketch Engine. German Web 2018: urshany.ch)<br>Frase preposicional *von*: *Diskussion* ***von Umsetzungsmöglichkeiten*** (Sketch Engine. German Web 2018: tanztherapie.de) |
| **Complemento prepositivo**<br>'aquele/aquilo não afetado: tema'<br>[situação] | *darüber* + oração subordinada completiva ou interrogativa: *rege Diskussionen* ***darüber, wie Verwaltung und Politik enger zusammenarbeiten können*** (Sketch Engine. German Web 2018: sozialestadt.at)<br>*darum* + oração subordinada completiva ou interrogativa: *Diskussion* ***darum, was das Besondere der Universitäten ist*** (Sketch Engine. German Web 2018: uni-kiel.de) |

*Nota.* A estrutura argumental proposta para o nome *Diskussion* foi realizada através do recurso fundamental ao dicionário PORTLEX e ao dicionário de Sommerfeldt e Schreiber (1983b), para além da consulta de *corpora* em Sketch Engine.

Analisando as diferentes realizações formais que se apresentam para os argumentos que costumam ocorrer na vizinhança do predicado nominal *Diskussion* (*vd*. tabela 5 *supra*), observa-se que algumas destas estruturas podem ser consideradas como ambíguas, no sentido já esclarecido em § 6.1. Nesta linha, apresentar-se-ão a seguir as realizações sintáticas onde a componente semântica representa a única possibilidade de diferenciação entre argumentos:

- Genitivo: a estrutura em genitivo ocorrendo à direita do predicado nominal *Diskussion*, considerando o quadro de posições sintáticas, pode desempenhar a função de complemento sujeito ('aquele/aquilo que realiza a ação') ou a função de complemento objeto ('aquele/aquilo não afetado: tema'). O recurso a uma consulta CQL em Sketch Engine, `[lemma="Diskussion"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"][tag="N.*"]`, evidencia a dificuldade que acarreta a falta de anotação semântica, pois os resultados extraídos podem corresponder com os esquemas *discussão de* [[humano]] ou *discussão de* [[imaterial]], para além de todas as outras possibilidades não actanciais que se listam em Sketch Engine. Alguns exemplos dos resultados são os seguintes:
    - *In der* ***Diskussion der Resultate*** *müssen die spezifischen Interpretationen der Studienergebnisse dargestellt werden*. (Sketch Engine. German Web 2018: henet.ch)

- *Menschenrechte entstehen aus Vernunftgründen, aus Wissen über den Menschen, aus der **Diskussion der Bürger**, aus der Auseinandersetzung mit der Realität.* (Sketch Engine. German Web 2018: politonline.ch)
- *Betrachtet man die öffentliche **Diskussion der Gegenwart**, so scheint es, dass die Werbung in Deutschland und zunehmend auch bei der EU-Kommission „schlechte Karten" hat*. (Sketch Engine. German Web 2018: theeuropean.de)

- Frase preposicional *von*: de acordo com Engel (2004), uma estrutura linguística encabeçada pela preposição *von* pode funcionar como agente ou complemento sujeito. Não obstante, trata-se de uma realização que dicionários valenciais como o de Sommerfeldt e Schreiber (1983b) ainda não recolhem. Por sua parte, no PORTLEX já aparece como possibilidade formal para a expressão do complemento sujeito e do complemento prepositivo ('aquele/aquilo não afetado: tema'). O recurso à preposição *von* em alemão para a expressão de mais de um complemento torna-se logo evidente através da utilização de uma consulta CQL em Sketch Engine como a que segue: `[lemma="Diskussion"][lemma="von"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Dat.*"]?[tag="N.*"].`
  - *Diese Domain befindet sich im Aufbau und ist aus der **Diskussion von Betroffenen** und Interessierten entstanden*. (Sketch Engine. German Web 2018: carmen-m.de)
  - *Wichtig ist Ihnen der konstruktive Austausch von Sichtweisen und die **Diskussion von Lösungsansätzen***. (Sketch Engine. German Web 2018: karriere.at)
  - *Anmoderiert wurde die **Diskussion von Manfred** Scholl und Stefanie Nejedlo*. (Sketch Engine. German Web 2018: manuel-westphal.de)
- Composto: o caso dos compostos, como já se referiu em § 6.1., destaca-se como sendo muito frequente pela facilidade de lexicalização que os compostos têm na língua alemã. O caso de *Podiumsdiskussion*, o mais frequente no *corpus* de trabalho em Sketch Engine, aparece já lematizado no sistema lexical OWID. Não obstante, analisando as partes dos compostos pode-se amiúde entender as estruturas resultantes como derivadas originariamente do esquema actancial. Com a execução de uma consulta CQL como `[lemma=".*diskussion"]`, obtemos exemplos como os elencados a seguir:

- *Ich wünsche uns allen eine spannende **Podiumsdiskussion** und gebe das Wort weiter*. (Sketch Engine. German Web 2018: bofo-ev.de)
- *Meiner Meinung nach ist diese **Systemdiskussion** völlig überschätzt*. (Sketch Engine. German Web 2018: dasgelbeblatt.de)
- *Die AbsolventInnen werden befähigt, eigenständige Beiträge zu germanistischen **Fachdiskussionen** zu leisten*. (Sketch Engine. German Web 2018: studium.at)

- Adjetivo: nesta dissertação defende-se a tese de que os adjetivos, para além de serem modificadores na frase nominal (cf. Arias, 2020), também podem agir como actantes e complementos do ponto de vista da gramática de valências. Torna-se necessário realizar uma consulta CQL com a demarcação do adjetivo como aparecendo obrigatoriamente em posição atributiva (`[tag="ADJA.*"][lemma="Diskussion"]`), já que se trata do caso marcado para os adjetivos preencherem *slots* actanciais:
  - *Bei der **anschließenden Diskussion** wurden die folgenden Aspekte formuliert*. (Sketch Engine. German Web 2018: zadoco.site)
  - *Carla war mit **politischen Diskussionen** beim Frühstück aufgewachsen*. (Sketch Engine. German Web 2018: weltbild.de)
  - *Es werden schon **rege Diskussionen** geführt und wichtige Infos ausgetauscht*. (Sketch Engine. German Web 2018: opelz-blog.de)

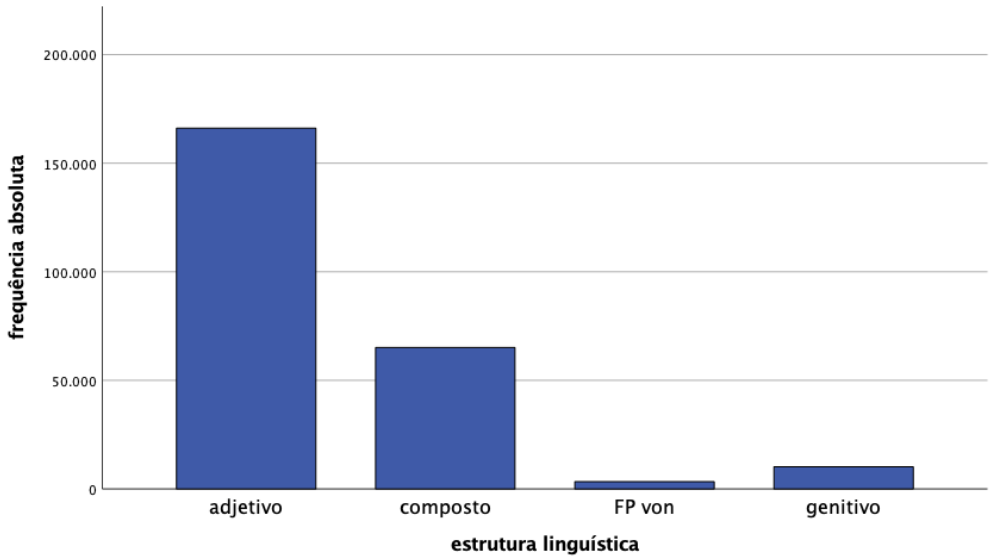
Estas estruturas podem, portanto, ser consideradas como ambíguas, de acordo com o defendido por Móia (2016, pp. 309-310), pois a presença do caráter ambíguo cinge-se à estrutura dos constituintes sintáticos. A única possibilidade de desambiguação reside, em consequência, no nível semântico. Como já se referiu ao longo deste trabalho, esta desambiguação constitui um dos pontos fortes da presente pesquisa.

O critério quantitativo e de frequência baseado no *corpus* de trabalho de Sketch Engine não permite retirar conclusões decisivas quanto à conformação dos esquemas actanciais, mas uma análise da frequência em termos absolutos da presença das diferentes estruturas listadas permite-nos deduzir qual é a sua representação em textos em língua alemã. Daí, tal como mostra a figura 11, deduz-se que o adjetivo em posição atributiva, assim como o composto, são as estruturas que ocorrem com maior frequência no *corpus* selecionado.

Aliás, deve-se esclarecer que quer os adjetivos quer os compostos acarretam maiores dificuldades na altura da desambiguação. Por um lado, os adjetivos podem desempenhar várias funções, entre as que se destacam a de modificadores (cf. Arias, 2020), a de colocativos (cf. Mel'čuk, 2015) e a de actantes (cf. Domínguez, 2011). Por outro lado, os compostos, para além de estarem mais lexicalizados internamente na língua, envolvem a aplicação de outros recursos do ponto de vista computacional, pois cumpre analisar a estrutura interna antes de poder anotá-los semanticamente. A partir do exposto, infere-se que a estrutura em genitivo era a mais adequada (em termos de frequência e de realização formal) para ser analisada na presente dissertação e isto será o que nos ocupará ao longo deste subcapítulo.

**Figura 11**

*Frequência absoluta em Sketch Engine: German Web 2018 das estruturas selecionadas*




Assim, e de acordo com o previamente mencionado, sabe-se que o genitivo pode ser utilizado para a expressão do complemento sujeito ('aquele/aquilo que realiza a ação') como para a expressão do complemento objeto ('aquele/aquilo não afetado: tema'). A aplicação do *script* desenhado *ad hoc* e baseado no pacote lexical {animado, humano} permitir-nos-á estudar a ocorrência de estruturas como *Diskussion der/des* [[humano]]. Não obstante, tal como afirmam Domínguez *et al*. (2018), a frequência em *corpus* nem sempre ajuda para a análise destas estruturas linguísticas e, posteriormente, para o desenvolvimento de aplicações de género lexicográfico.

A tese defendida por Domínguez *et al*. (2018) é facilmente verificada através da observação da tabela 6 *infra*, pois descobre-se logo que apenas 15% das estruturas listadas corresponde com o padrão *Diskussion der/des* [[humano]], sendo que as outras costumam apresentar a estrutura *Diskussion*

*der/des* [[imaterial]]. Destarte, afirma-se que apenas nas frases nominais *Diskussion der Teilnehmer*, *Diskussion des Forums* e *Diskussion dieser Nutzer*, a estrutura em genitivo pode ser interpretada como complemento sujeito ('aquele/aquilo que realiza a ação').

**Tabela 6**

*Dados mais frequentes para a estrutura* Diskussion *+ genitivo em Sketch Engine: German Web 2018*

| lema | frequência |
|---|---|
| Diskussion die Ergebnis | 471 |
| Diskussion die Thema | 269 |
| Diskussion die Frage | 151 |
| Diskussion die Inhalt | 130 |
| Diskussion diese Frage | 115 |
| Diskussion diese Thema | 113 |
| Diskussion die Teilnehmer | 90 |
| Diskussion die Begriff | 79 |
| Diskussion die Forum | 71 |
| Diskussion die Problem | 66 |
| Diskussion diese Nutzer | 64 |
| Diskussion diese Art | 61 |
| Diskussion die Konzept | 54 |
| Diskussion die Entwurf | 50 |
| Diskussion die Verhältnis | 44 |
| Diskussion die Beitrag | 40 |
| Diskussion die Text | 39 |
| Diskussion die Zeit | 37 |
| Diskussion die These | 35 |
| Diskussion die Möglichkeit | 35 |
| Diskussion die Bericht | 34 |

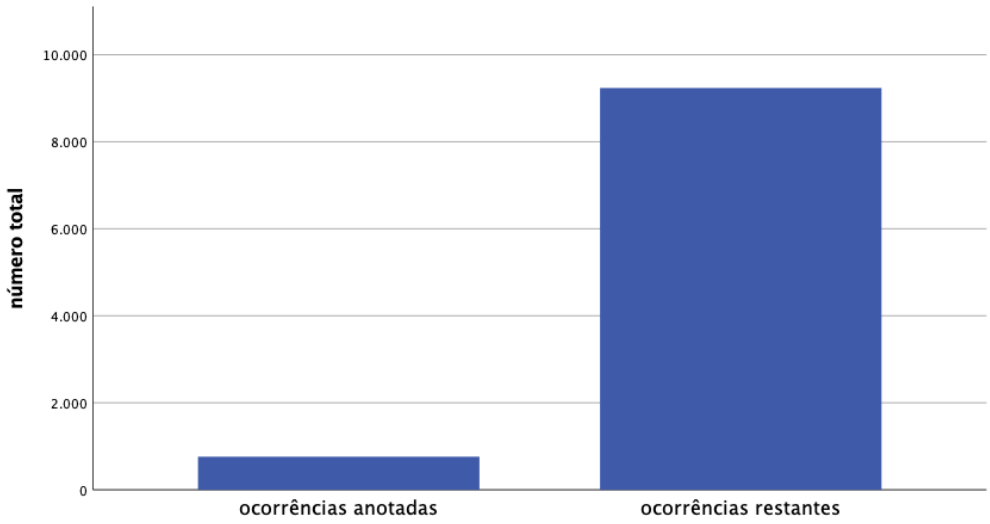
Com o recurso a parâmetros de frequência, apreciam-se três possibilidades actanciais para a unidade lexical *Diskussion* combinada com o genitivo. A primeira delas foi já explicada *supra* e faz referência ao complemento sujeito quando em frases como *Diskussion der Teilnehmer*. A segunda delas refere-se ao complemento objeto, em que através da estrutura em genitivo se transmite informação sobre o tema da discussão (*Diskussion der Ergebnisse* ou *Diskussion des Themas*). Por último, a terceira hipótese tem a ver com a realização classificativa (cf. Domínguez, 2014b; Engel, 2004), com a presença da frase nominal *Diskussion dieser Art*. Todavia, não parece relevante fornecer aqui mais informação

sobre as diferentes possibilidades de combinação livre e argumental, uma vez que o objetivo principal desta dissertação é avaliar o funcionamento do *script* desenhado.

Para tal propósito, extraíram-se (seguindo o fluxo de trabalho realizado em § 6.1.) 10000 concordâncias do *corpus* selecionado em Sketch Engine aplicando a consulta CQL *supra* referida. Todas as ocorrências, que correspondem ao esquema *Diskussion* + genitivo, foram analisadas com o *script* desenhado no âmbito deste trabalho partindo do pacote lexical {animado, humano}. Desta forma, anotou-se automaticamente 8,26% das concordâncias com o *tag* semântico `<sem_tag type="human">`. Não obstante, esta informação em termos de frequência absoluta não se pode considerar relevante *stricto sensu* para o objetivo desta dissertação, pois é possível que a estrutura *Diskussion der/des* [[humano]] apareça num número relativamente reduzido de ocasiões. A figura 12 *infra* mostra somente esta percentagem, sendo que as ocorrências restantes correspondem com outras estruturas argumentais sem o traço {animado, humano}, caso o *script* esteja a funcionar de acordo com o pré-estabelecido.

**Figura 12**

*Frequência absoluta das ocorrências anotadas como {animado, humano} para* Diskussion




Contudo, este passo é essencial para facilitar a manipulação posterior dos dados, uma vez que se conseguem extrair facilmente as estruturas já anotadas com o traço {animado, humano}, para depois ponderar os resultados atingidos. Como já se referiu, o pacote lexical de trabalho inclui unidades lexicais com o traço semântico [instituição], para além do puramente [humano]. Aliás, dicionários de valências como o PORTLEX ou Sommerfeldt e Schreiber (1983b) afirmam que para o *slot* semântico do

complemento sujeito podem ocorrer ambos os traços mencionados. As concordâncias anotadas que aparecem mais frequentemente são as seguintes[51]:

```
86 <kwic>Diskussion der <sem_tag type="human">Teilnehmer</sem_tag></kwic>
61 <kwic>Diskussionen dieses <sem_tag type="human">Nutzers</sem_tag></kwic>
25 <kwic>Diskussion der <sem_tag type="human">Mitglieder</sem_tag></kwic>
13 <kwic>Diskussion der <sem_tag type="human">Jury</sem_tag></kwic>
```

A partir das ocorrências supra apresentadas, deduz-se que o *script* consegue identificar os elementos lexicais que foram incluídos anteriormente no pacote lexical e consegue, aliás, anotá-los semanticamente. Esta hipótese já fora verificada para o substantivo *Bericht*. Alguns problemas apresentados já em § 6.1. mostram-se constantes e persistem com o núcleo nominal *Diskussion*. Tal é o caso da anotação dupla em compostos hifenizados, pois ambos os constituintes destes compostos foram introduzidos no pacote lexical. Em consequência, a máquina reconhece-os como pertencentes ao pacote {animado, humano} e coloca a etiqueta semântica correspondente:

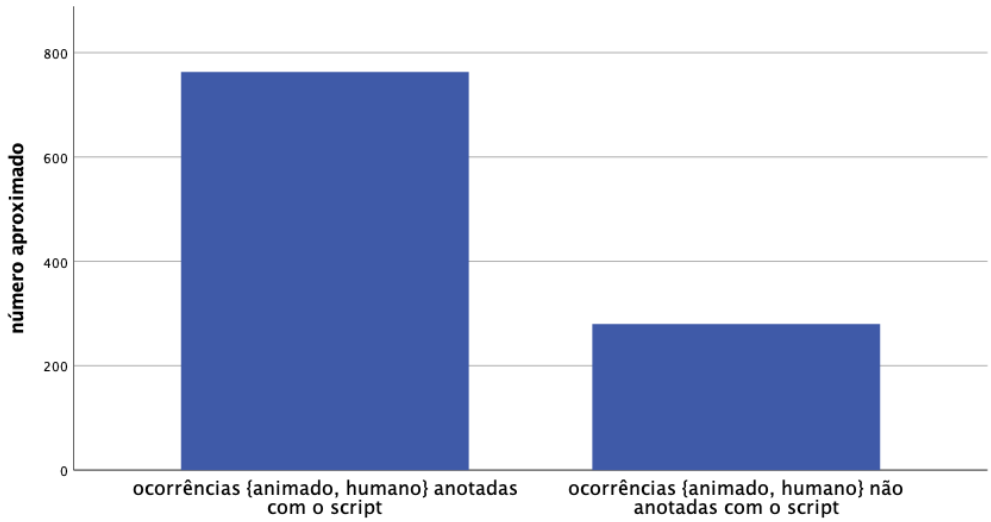
```
<kwic>Diskussionen der <sem_tag type="human">Bund</sem_tag>-Länder-<sem_tag type="human">Kommission</sem_tag></kwic>
<kwic>Diskussionen der <sem_tag type="human">SPD</sem_tag>-<sem_tag type="human">Fraktion</sem_tag></kwic>
```

Para avaliar o adequado funcionamento do script desenvolvido, deve-se estudar se existem concordâncias que num procedimento manual seriam etiquetadas como {animado, humano} e que a máquina não reconheceu. Neste sentido, o recurso a expressões regulares tornou-se fundamental na presente pesquisa, pois trata-se de uma forma mesurada para descobrir, numa abordagem (semi)automática, que entidades não são anotadas. Destarte, conseguem-se também vaticinar os possíveis motivos.

[51] Como já se mencionou antes, o número à esquerda corresponde à frequência absoluta no ficheiro com anotação semântica.

**Figura 13**

*Número aproximado de ocorrências {animado, humano} no corpus*




*Nota.* O número de ocorrências anotadas automaticamente com o *script* foi calculado também de forma automática. No entanto, as ocorrências não anotadas que podem ser classificadas como {animado, humano} foram contadas manualmente com a execução de expressões regulares.

A figura 13 *supra* ilustra a frequência de ocorrências que, através da aplicação de métodos semiautomáticos, podem ser classificadas também como {animado, humano}. Porém, observa-se que aproximadamente 27% das concordâncias correspondentes com o pacote lexical selecionado não foram anotadas de forma automática. Assim sendo, consegue-se apresentar uma relação de elementos não etiquetados diretamente com o *script*, o que permite estabelecer hipóteses sobre os motivos que impedem essa anotação. Consequentemente, o *script* poderá ser retroalimentado com a informação derivada desta análise, de modo a tornar-se mais preciso.

- Ausência de algumas unidades lexicais no pacote lexical de trabalho: um dos motivos pelos quais algumas das entidades não foram reconhecidas automaticamente como {animado, humano} tem a ver com o facto de o pacote lexical ser limitado em termos de vocabulário. Isto acontece fundamentalmente com duas classes de entidades lexicais. A primeira diz respeito aos compostos, pelo que a execução de expressões regulares facilita o procedimento. Desta forma, a utilização, por exemplo, da expressão `.*leute` permite-nos encontrar ocorrências como *Diskussion der Fachleute* ou *Diskussion der Vertrauensleute*. A segunda corresponde sobretudo com a presença de palavras morfossintáticas em feminino plural que não foram recolhidas no nosso pacote lexical.

Assim, a expressão `.*innen</kwic>` faculta a anotação de estruturas como *Diskussion der Nutzerinnen* ou *Diskussionen der Campteilnehmerinnen*.

- Presença de siglas que funcionam como complemento sujeito: as siglas, normalmente para representarem grupos, associações ou instituições, aparecem em 1,4% das ocorrências totais analisadas para o núcleo nominal *Diskussion*. Contudo, o *script* não as reconhece como {animado, humano} e estão, em consequência, desprovidas de etiquetagem semântica. Trata-se de casos como, por exemplo, *Diskussion der AG* ou *Diskussionen der AfD*.
- Presença de particípios que agem como complemento sujeito: um problema que conduz à falta de anotação semântica em ocasiões prende-se com a ocorrência de particípios I e II em alemão para o *slot* sintático-semântico do complemento sujeito. Assim, e de modo a ilustrar este problema, para o caso do particípio I, a execução da expressão regular `.*nden</kwic>$`, permite-nos extrair de forma semiautomática ocorrências como *Diskussion der Anwesenden* ou *Diskussion der Teilnehmenden*, que deveriam ser anotadas com a etiqueta semântica {animado, humano}.
- Presença de estruturas incompletas ou inadequadamente processadas: trata-se de casos em que existe ruido do ponto de vista computacional, sendo que a máquina não reconhece as entidades lexicais pela existência de estruturas desconhecidas. A seguinte concordância deveria, por exemplo, ser eliminada numa fase de pré-processamento para evitar a sua incorporação no *corpus* de trabalho, uma vez que leva a problemas ao tentarmos anotar a estrutura: `<kwic>Diskussion d...</kwic>`.

Alguns casos como os três primeiros elencados nesta lista podem ser facilmente resolvidos em fases posteriores do trabalho, uma vez que só implicariam acrescentar as unidades lexicais encontradas e não anotadas ao pacote lexical original {animado, humano}. Isto não aponta, portanto, para um funcionamento incorreto do *script*, mas apenas evidencia que o pacote de lexical é uma lista de vocabulário finita, cuja composição foi realizada, pelo menos parcialmente, seguindo critérios manuais e de análise (*vd*. capítulo 5.2.).

Torna-se, contudo, imprescindível mencionar que existem outros aspetos que fundamentam o funcionamento do *script* consoante os objetivos pretendidos na presente dissertação:

- Correspondência de unidades não anotadas com outras estruturas argumentais: no presente capítulo, visava-se apresentar os resultados para a anotação da estrutura em

genitivo com o traço semântico {animado, humano}. Os esquemas actanciais que ficam fora dessa caracterização não devem ser anotados pelo *script* salvo que ele não esteja a operar adequadamente. Uma análise das estruturas não anotadas, permite-nos corroborar o funcionamento adequado do *script*. O exemplo mais evidente é a estrutura *Diskussion der Ergebnisse*, que aparece em 477 concordâncias, ou *Diskussion des Themas*, cuja frequência em termos absolutos no nosso *corpus* é de 312 ocorrências. Em suma, estabelece-se que estas estruturas, correspondentes ao esquema *Diskussion der/des* [[imaterial]] não são anotadas porque não correspondem com o traço ontológico {animado, humano}.

- Correspondência de unidades não anotadas com realizações de adjuntos ou circunstantes: é também frequente encontrarmos estruturas como *Diskussion der Zeit* ou *Diskussion des Jahres*, onde a estrutura em genitivo não desempenha a função de um actante. Neste caso, como no anterior, deve-se destacar que a falta de anotação é a prova do funcionamento apropriado do *script*.

Em síntese, como já foi salientado em relação a *Bericht*, corrobora-se novamente que o *script* desenhado *ad hoc* está a cumprir a função principal para a que foi criado, sendo que possibilita uma manipulação posterior de dados mais fácil tanto para o lexicógrafo ou linguista que se ocupa da sua análise, como para a máquina que lerá os dados lexicais introduzidos. Por último, também se afirma que os casos em que o *script* não anota unidades lexicais que pertencem ao pacote {animado, humano} podem vir a ser reduzidas com a introdução daquelas entidades selecionadas na análise semiautomática e manual no pacote lexical original, de modo a alargá-lo e enriquecê-lo.

### 6.3. O substantivo *Frage*

Na informação lexicográfica atinente ao substantivo *Frage* e incluída nos dicionários consultados no âmbito da presente pesquisa (DWDS, n. d.; DUDEN, n. d.; Sommerfeldt & Schreiber, 1983b), observa-se que se trata de um nome polissémico, sendo que no caso da primeira aceção ('pergunta'), a unidade lexical *Frage* pode fazer parte do campo lexical da comunicação, mas no segundo caso, a paráfrase do significado remete para outro sentido diferente ('problema, assunto'):

„1. mündliche oder auch schriftliche Äußerung, mit der sich jmd. an jmdn. wendet, um etw. von ihm zu erfahren

2. Angelegenheit, die eine Erörterung, Klärung, Entscheidung verlangt; Problem, Sache, Angelegenheit"

(DWDS, n. d.)

„1. eine Antwort, Auskunft, Erklärung, Entscheidung o. Ä. fordernde Äußerung, mit der sich jemand an jemanden wendet

2. Problem; zu erörterndes Thema, zu klärende Sache, Angelegenheit“ (DUDEN, n. d.)

„1. Äußerung an jmdn., um etwas zu erfahren

2. Problem“ (Sommerfeldt & Schreiber, 1983b, p. 194)

Para além da definição, que oferece apenas informação sobre o significado associado a um significante determinado, no artigo lexicográfico incluem-se outros dados lexicográficos que podem ser interessantes para o objetivo da presente dissertação. Fala-se, nomeadamente, em exemplos, pois eles podem auxiliar a função de produção ou codificação linguística (cf. Fuertes-Olivera & Bergenholtz, 2018, p. 272).

Em primeiro lugar, o DWDS (n. d.) oferece informação implícita sobre algumas possibilidades de realização formal para a estrutura argumental do substantivo *Frage*, embora os dados lexicográficos não sejam tratados de forma pormenorizada. Desta forma, associados à primeira aceção (‘pergunta’), contamos com exemplos como *eine Frage an mich* ou *Frage zur Person*, que podem considerar-se como relevantes para o padrão actancial por providenciarem informação sobre dois argumentos prepositivos diferentes. Não obstante, os utilizadores não obtêm explicações relativas ao uso dessas estruturas linguísticas. Por sua parte, na segunda aceção (‘problema’), aparecem exemplos com frases em genitivo (*die Frage der Groß- und Kleinschreibung*) ou até adjetivos (*die nationale, soziale Frage*) como complementos, mesmo que a informação sobre o seu uso não apareça explicitada.

Em segundo lugar, no DUDEN (n. d.) apresenta-se também informação não explícita ligada à estrutura argumental nominal. Assim, a respeito da paráfrase de significado concernente à proposição semântica ‘resposta’, aparece apenas o exemplo *Fragen zur Person und zur Sache*, pois os outros exemplos inseridos têm a ver com a recção do verbo e não do substantivo *per se* (vejam-se, por exemplo, os casos de *an jemanden eine Frage richten* ou *auf eine Frage antworten*). No tocante à segunda aceção (‘problema’), providencia-se um maior número de exemplos no âmbito do esquema actancial nominal, uma vez que o utente obtém informação de forma passiva sobre, pelo menos, dois argumentos: o complemento prepositivo com o conteúdo semântico ‘aquele/aquilo não afetado: tema’ (*die Frage nach dem Sinn des Lebens*) e o complemento objeto com o mesmo traço semântico (*eine Frage der Ehre* ou *eine Frage des Geldes*).

Em síntese, deduz-se que os dicionários gerais de língua alemã oferecem escassa informação sobre a estrutura argumental do substantivo *Frage*, pois estes dados lexicográficos relativos ao esquema

actancial linguístico são apenas incluídos na secção dos exemplos dentro do artigo lexicográfico. De modo a sopesarmos esta falta de descrição da estrutura argumental em dicionários de língua geral, e consoante as descrições oferecidas no dicionário PORTLEX e por Sommerfeldt e Schreiber (1983b), a tabela 7 *infra*, visa elencar os argumentos que aparecem com este substantivo partindo fundamentalmente da primeira aceção ('pergunta'), por ser a correspondente ao campo lexical da comunicação.

**Tabela 7**

*Estrutura argumental do nome* Frage

| | |
|---|---|
| **Complemento sujeito**<br>'aquele/aquilo que realiza a ação'<br>[humano] [instituição] | Genitivo: *Fragen* ***der Zuhörer*** (Sketch Engine. German Web 2018: elisabeth-buechle.de)<br>Frase preposicional *von*: *Fragen* ***von Studierenden*** (Sketch Engine. German Web 2018: solarportal24.de)<br>Determinante possessivo: ***unsere*** *Frage* (Sketch Engine. German Web 2018: bifo.de) |
| **Complemento objeto**[52]<br>'aquele/aquilo não afetado: tema'<br>[imaterial] | Genitivo: *Fragen* ***des Lebens*** (Sketch Engine. German Web 2018: hoppsaala.de)<br>Palavra composta: ***Sicherheits****frage* (Sketch Engine. German Web 2018: fifaplanet.de)<br>Adjetivo: ***soziale*** *Frage* (Sketch Engine. German Web 2018: npd-oberland.de) |
| **Complemento prepositivo**<br>'aquele/aquilo não afetado'<br>[humano] [instituição] | Frase preposicional *an*: *die Frage* ***an den Kandidaten*** (Sketch Engine. German Web 2018: lehmann-coll.de) |
| **Complemento prepositivo**<br>'aquele/aquilo não afetado: tema'<br>[imaterial] | Frase preposicional *nach*: *die Frage* ***nach dem Ziel*** (Sketch Engine. German Web 2018: stahlaufstahl.de)<br>Frase preposicional *über*: *eine Frage* ***über die Bewerbung*** *für eine Praktikumsstelle* (Sketch Engine. German Web 2018: abi.de) |

[52] Note-se que o complemento objeto costuma estar associado à segunda aceção de *Frage* ('problema') (cf. Sommerfeldt & Schreiber, 1983b).

| | |
|---|---|
| | Frase preposional *zu*: *eine Frage* ***zu einem Produkt*** (Sketch Engine. German Web 2018: hochdruckreinigerzubehoer.ch)<br><br>Frase preposicional *von*: *Frage* ***von der Sprachwahl*** (Sketch Engine. German Web 2018: taz.de) |
| **Complemento prepositivo**<br><br>'aquele/aquilo não afetado: tema'<br><br>[situação] | *danach* + oração subordinada completiva ou interrogativa: *Frage* ***danach, worum es der anderen Person eigentlich geht*** (Sketch Engine. German Web 2018: lubbers.at)<br><br>*darüber* + oração subordinada completiva ou interrogativa: *Frage* ***darüber, wie diese Beziehung zustande kommt*** (Sketch Engine. German Web 2018: t-online.de)<br><br>*dazu* + oração subordinada completiva ou interrogativa: *Frage* ***dazu, wem der Produktivitätsfortschritt und dessen Ergebnisse zugutekommen*** (Sketch Engine. German Web 2018: freitag.de) |

*Nota.* A estrutura argumental proposta para o nome *Frage* foi realizada através do recurso fundamental ao dicionário PORTLEX e ao dicionário de Sommerfeldt e Schreiber (1983b), para além da consulta de *corpora* em Sketch Engine.

Infere-se, portanto, que existem várias realizações formais para os argumentos que se incluem e registam como sendo complementos do predicado nominal *Frage*. Aliás, alguns argumentos podem ser expressos através do recurso a fórmulas linguísticas diferentes e daí surge a necessidade de desambiguação. Tal como se realizou em § 6.1. e § 6.2., partir-se-á no presente subcapítulo das quatro estruturas linguísticas já analisadas para os outros nomes, embora no caso de *Frage* cumpra destacar outros aspetos:

- Genitivo: a realização formal em genitivo é empregue para preencher o *slot* semântico-sintático pertencente ao complemento sujeito ('aquele/aquilo que realiza a ação') e também para o complemento objeto ('aquele/aquilo não afetado: tema'). Com a execução de uma consulta CQL como a seguinte `[lemma="Frage"][tag="(ART\.(Def|Indef)|PRO.(Dem|Poss).Attr).Gen.*"] [tag="N.*"]` nem sempre podemos descobrir de que complemento se trata, uma vez que a expressão formal é idêntica em todos os casos. Assim, o nível de abstração que comporta trabalhar com frases nominais acaba por impedir em ocasiões que saibamos se estamos a trabalhar

com unidade lexical *Frage* com o significado 'pergunta' ou 'assunto, problema'. Isto aprecia-se claramente em frases como *Frage der Gerechtigkeit*, pois pode significar (sem contexto) 'pergunta de justiça' ou 'assunto/problema de justiça'. Só uma anotação semântica ia permitir esta desambiguação, pois a consulta CQL só nos permite peneirar as ocorrências com uma determinada estrutura sintática ou formal:

- *Aber Anregungen und **Fragen der Mitglieder** können über diesen Weg natürlich der Naturschutzbehörde mitgeteilt und im Naturschutzbeirat beraten werden*. (Sketch Engine. German Web 2018: und-rlp.de)
- *In den letzten Jahren standen vor allem **Fragen der Gendergerechtigkeit** oder der Kampf gegen Rassismus und Rechtsextremismus im Vordergrund.* (Sketch Engine. German Web 2018: junge-gruene.at)
- *Das Projekt untersucht hierbei **Fragen der Grenzziehung** und des Minderheitenschutzes ebenso wie der (Zwangs-)Migration.* (Sketch Engine. German Web 2018: uni-due.de)

- Frase preposicional *von*: a frase preposicional com *von* substitui frequentemente, de acordo com Engel (2004), a estrutura em genitivo para a transmissão do traço semântico 'aquele/aquilo que realiza a ação'. No caso do predicado nominal *Frage*, a frase preposicional com *von* seguida de um substantivo [humano] costuma corresponder com o padrão *Frage von* [[humano]] ('a pergunta de [[humano]]'). Porém, a ocorrência do nome *Frage* acompanhado de uma frase preposicional *von* e um substantivo não humano (com o traço [imaterial], por exemplo), não nos permite sempre esclarecer se se trata de uma aceção ou outra ('pergunta' ou 'assunto'). Isto observa-se, aliás, com a aplicação de consultas CQL como `[lemma="Frage"] [lemma="von"] [tag="(ART\.(Def|Indef)| PRO. (Dem|Poss).Attr) .Dat.*"]?[tag="N.*"]`:
  - *Die **Frage von Waffenlieferungen** müsse immer vor dem Hintergrund der Stabilisierung im Nahen Osten diskutiert werden*. (Sketch Engine. German Web 2018: nahost-politik.de)
  - *Immer wieder kommen **Fragen von Leuten**, die auf unsere Internetseite gestoßen sind und den VW166 Schwimmwagen so toll finden*. (Sketch Engine. German Web 2018: wh-schwimmwagen.de)

- *Zwar steht Deutschland in **Fragen von Mülltrennung**, Energieeffizienz und Wohnraumisolierung tatsächlich ganz weit oben im internationalen Vergleich*. (Sketch Engine. German Web 2018: theeuropean.de)

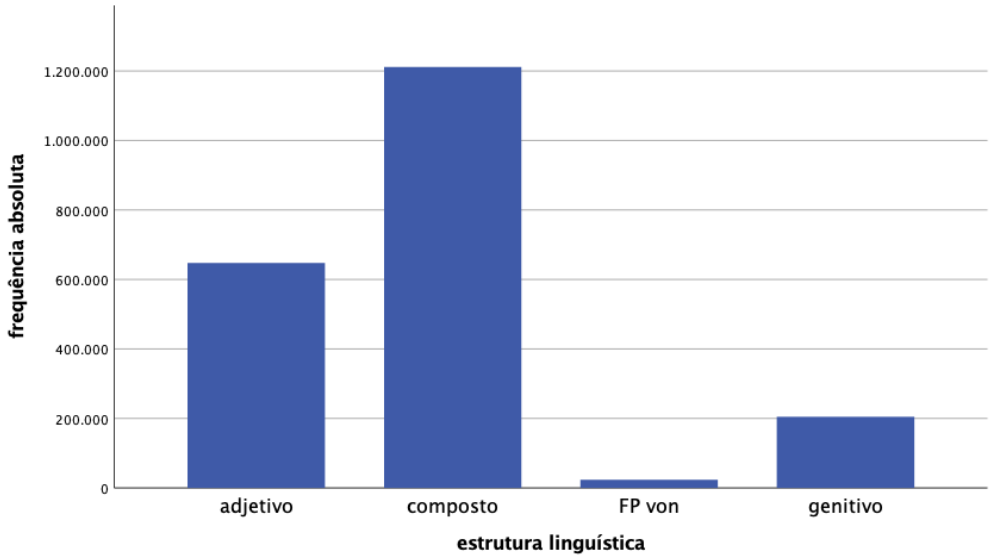
- Composto: através da consulta CQL `[lemma=.*frage"]` em Sketch Engine não só obtemos, no caso de *Frage*, compostos formais, como também se fornecem outras unidades lexicais plenamente lexicalizadas, como é o caso de *Umfrage* ou *Nachfrage*. Isto faz com que a frequência absoluta de aparição no *corpus* aumente de forma notória. Vê-se, em consequência, que os resultados da mencionada consulta podem ser muito diferentes:
  - *Diese **Nachfrage** wird in den nächsten Jahrzehnten noch weiter stark steigen*. (Sketch Engine. German Web 2018: db.com)
  - *Wenn du so oft Chancen liegen lässt, ist das irgendwann eben auch eine **Qualitätsfrage***. (Sketch Engine. German Web 2018: kreis-anzeiger.de)
  - *Wir stehen für Erfahrung und Kompetenz in allen **Gesundheitsfragen***. (Sketch Engine. German Web 2018: sonnen-apotheke.de)
- Adjetivo: embora consideremos os adjetivos como sendo realizações formais possíveis no âmbito da estrutura argumental, as frases adjetivas não foram incluídas nesta ocasião na descrição realizada na tabela 7 pelo facto de contarem com maior grau de ambiguidade. Arias (2020, p. 31) destaca que frases como *gerichtliche Frage* podem contar com várias interpretações: 'a pergunta do tribunal', 'a pergunta de tipo jurídico', 'o assunto de jurisdição', entre outras. Para além disto, não se deve esquecer que os adjetivos podem funcionar apenas como modificadores (cf. Arias, 2020) ou também como colocativos (cf. Mel'čuk, 2015). Esta dificuldade, assim como a heterogeneidade dos exemplos obtidos com a consulta CQL `[tag="ADJA.*"][lemma="Frage"]` fez com que a descrição dos adjetivos não fosse decisiva na presente pesquisa.
  - *Für **weitere Fragen** stehen wir gerne zur Verfügung*. (Sketch Engine. German Web 2018: kikaj.com)
  - ***Praxisorientierte Fragen** werden mit wissenschaftlichen Methoden bearbeitet*. (Sketch Engine. German Web 2018: bayern.de)
  - ***Sozialhistorische Fragen**, aber auch Fragen der Familienforscher, können in diesem Band Antwort finden*. (Sketch Engine. German Web 2018: heimatverein-viersen.de)

Em suma, estas estruturas revelam-se como ambíguas (no sentido de Móia, 2016), pois através da estrutura argumental pode-se ativar um significado ou outro. É por isso que se defende que uma anotação semântica de todas as estruturas poderia ajudar com a desambiguação, na medida em que existe consenso em que a estrutura *Frage der/des* [[humano]] transmite o significado de 'a pergunta de [[humano]]', sendo que a unidade lexical com o traço [humano] pode classificar-se como agente ou complemento sujeito e enquadrar-se no âmbito de aplicação da primeira aceção *supra* listada.

Deste modo, e seguindo a abordagem realizada em § 6.1. e § 6.2., veremos qual é a frequência absoluta das quatro estruturas linguísticas selecionadas no corpus de trabalho (*vd.* figura 14). Os compostos são a estrutura mais frequente, mas deve-se levar em consideração, como já se mencionou anteriormente, que não só são contabilizados os compostos *stricto sensu* neste caso, mas também são incluídas na contagem palavras derivadas por prefixação como *Umfrage* ou *Nachfrage*.

**Figura 14**

*Frequência absoluta em Sketch Engine: German Web 2018 das estruturas selecionadas*




À vista disto, escolhe-se a estrutura em genitivo para pôr em funcionamento o *script*, uma vez que foi igualmente verificado que a consulta CQL para este caso pode proporcionar dados linguísticos muito variados e até relativos a duas aceções diferentes. Assim, a aplicação do pacote lexical {animado, humano} permitir-nos-á filtrar as concordâncias selecionadas para extrair só as ocorrências que correspondem à estrutura argumental *Frage der/des* [[humano]].

Contudo, antes de executarmos o *script* para a anotação semântica das concordâncias retiradas do *corpus* em Sketch Engine, parece necessário fazermos um excurso às frases mais frequentes

com a estrutura em genitivo seguindo o predicado nominal *Frage*. Na tabela 8 *infra* elencam-se as 20 estruturas mais frequentes com a unidade lexical *Frage* e o genitivo.

**Tabela 8**

*Dados mais frequentes para a estrutura* Frage *+ genitivo em Sketch Engine: German Web 2018*

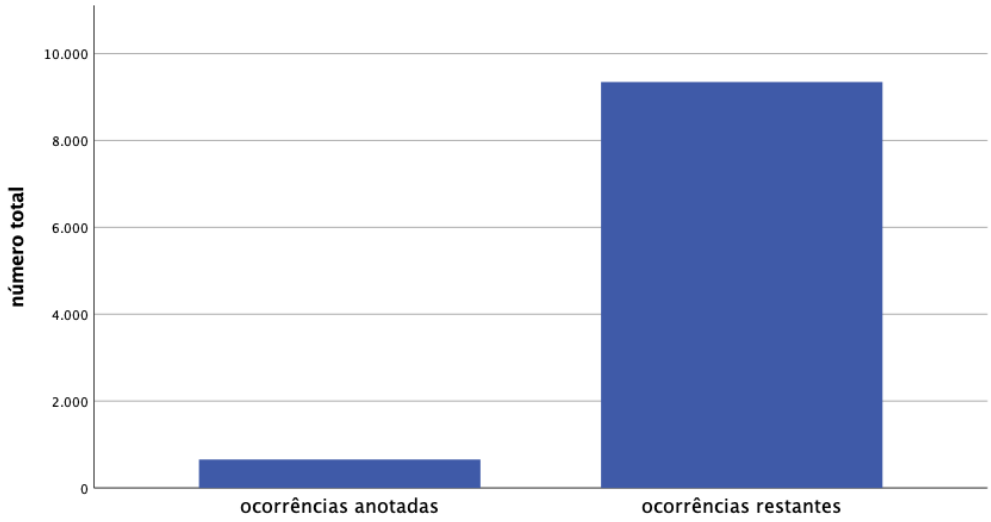
| **lema** | **frequência** |
|---|---|
| Frage die Zeit | 16432 |
| Frage die Leben | 1814 |
| Frage die Alter | 1145 |
| Frage die Geld | 1070 |
| Frage die Sicherheit | 1043 |
| Frage die Finanzierung | 1043 |
| Frage unsre Zeit | 980 |
| Frage die Zulässigkeit | 967 |
| Frage die Geschmack | 933 |
| Frage die Kind | 870 |
| Frage die Glaube | 868 |
| Frage die Teilnehmer | 708 |
| Frage die Perspektive | 696 |
| Frage die Besucher | 605 |
| Frage die Ehre | 597 |
| Frage die Preis | 589 |
| Frage die Schüler | 587 |
| Frage die Gerechtigkeit | 583 |
| Frage die Einzelfall | 579 |
| Frage die Haftung | 577 |
| Frage die Journalist | 566 |

Da análise dos dados linguísticos mais frequentes, deduz-se que 25% das estruturas selecionadas podem ser anotadas com o *tag* {animado, humano}, e nelas o padrão actancial corresponde com o traço semântico 'aquilo/aquele que realiza a ação'. Deve-se lembrar aqui que Domínguez *et al*. (2021) centravam o foco na necessidade de recorrer a outros parâmetros, para além da frequência, no âmbito de desenho de ferramentas alicerçadas na gramática valencial. A frequência não deve, de acordo com os autores, ser o único critério para filtrar ocorrências. É por isso que o *script* desenvolvido *ad hoc* nesta dissertação permitirá retirar conclusões mais definitivas sobre a combinação livre do lexema *Frage* com entidades que podem ser marcadas com o traço {animado, humano}.

A figura 15 *infra* mostra, em consequência, as vezes que o *script* etiquetou uma entidade com o *tag* {animado, humano}. Do total de 10000 concordâncias extraídas de Sketch Engine, somente 6,54% foram anotadas com essa etiqueta. Não obstante, esta informação só é relevante para sabermos quantas unidades podem ser extraídas semanticamente por contarem com o traço {animado, humano}. Assim sendo, não se pode concluir desta afirmação se o *script* está a mostrar um funcionamento adequado.

**Figura 15**

*Frequência absoluta das ocorrências anotadas como {animado, humano} para* Frage




Esta análise serve, todavia, como ponto de partida para a ulterior manipulação dos dados linguísticos já anotados com o *tag* {animado, humano}. Uma análise das entidades previamente etiquetadas permite-nos observar que as seguintes frases nominais são as mais frequentes entre as concordâncias com as quais se trabalha:

44 <kwic>Fragen der <sem_tag type="human">Besucher</sem_tag></kwic>

41 <kwic>Fragen der <sem_tag type="human">Teilnehmer</sem_tag></kwic>

36 <kwic>Fragen der <sem_tag type="human">Schüler</sem_tag></kwic>

32 <kwic>Fragen der <sem_tag type="human">Kinder</sem_tag></kwic>

Das ocorrências *supra* apresentadas, deduz-se que o *script* reconhece as unidades lexicais que podem ser anotadas com o *tag* e que acrescenta uma etiqueta XML para facilitar a posterior manipulação de dados lexicográficos. Tal como se explicou já na análise dos substantivos *Bericht* e *Diskussion*, o *script* anota algumas unidades de forma pouco adequada, uma vez que os compostos grafados com hífen são considerados pela máquina como duas palavras separadas. Destarte, aparecem frases nominais com

dupla anotação, uma para cada um dos elementos do composto, e outras com só um elemento anotado, como se mostra a seguir:

```
<kwic>Fragen der Energie-<sem_tag type="human">Fans</sem_tag></kwic>
```

A vantagem significativa que apresenta esta anotação "incorreta" é que permite marcar o segundo elemento de compostos pouco frequentes com o *tag* {animado, humano}, facilitando consequentemente a sua análise linguística e estatística. Não obstante, também aparecem outras ocorrências em que a anotação é completamente incorreta devido a erros prévios durante a fase de compilação de *corpus* e de pré-processamento. A presença de ruido pode levar a uma etiquetagem inapropriada:

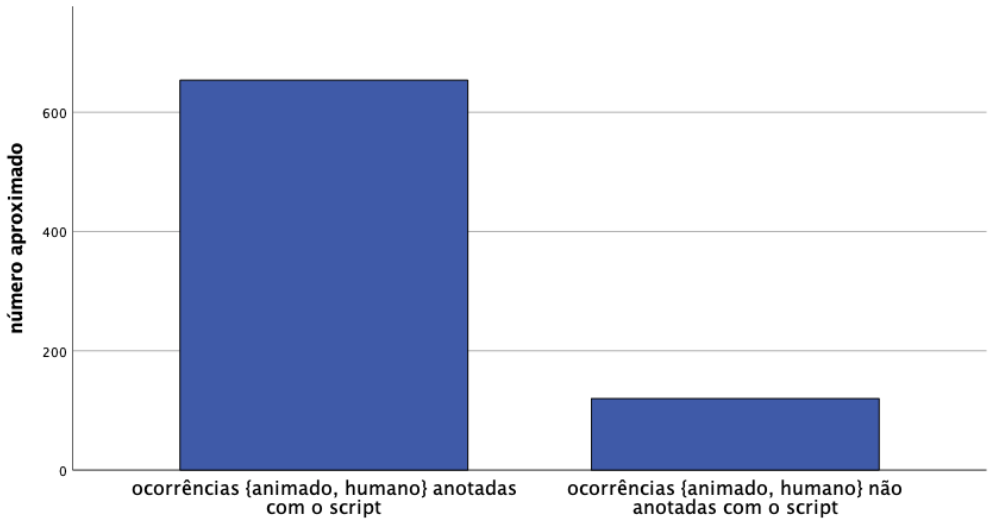
```
<kwic>Fragen des <sem_tag type="human">Personen</sem_tag>-und</kwic>
```

Neste caso concreto existem inclusive erros gramaticais (*Fragen *des Personen*), pelo que estas ocorrências deveriam ser eliminadas do *corpus* de trabalho para evitar a retirada de conclusões pouco fiáveis. Contudo, o facto de o *script* funcionar adequadamente é demonstrado com a sua capacidade para anotar as entidades que podem ser reconhecidas como {animado, humano}. As falhas que aparecem ligadas ao *script* têm como motivo, por vezes, a existência de erros provenientes da fase de pré-processamento textual.

Devemos ainda prestar atenção especial àqueles casos que são considerados, do ponto de vista da semântica categorial (Engel, 2004), como {animado, humano}, mas que não foram anotados automaticamente pelo *script*. A figura 16 *infra* mostra que, para o nome *Frage*, cerca de 84% das ocorrências que podem ser marcadas com esta etiqueta, foram encontradas e, consequentemente, anotadas pelo *script* de forma automática. Para a percentagem restante (16%, aproximadamente) de concordâncias não anotadas, uma aplicação de carácter semiautomático permite-nos deduzir alguns dos motivos.

**Figura 16**

*Número aproximado de ocorrências {animado, humano} no corpus*




*Nota.* O número de ocorrências anotadas automaticamente com o *script* foi calculado também de forma automática. No entanto, as ocorrências não anotadas que podem ser classificadas como {animado, humano} foram contadas manualmente com a execução de expressões regulares.

Desta forma, seguidamente apresentar-se-ão alguns dos elementos {animado, humano} que não foram anotados pelo *script*, visando oferecer uma justificação plausível. Tal como se procedeu para a análise dos nomes *Bericht* e *Diskussion*, o recurso a expressões regulares torna-se essencial para a manipulação de todos os dados linguísticos não anotados com que contamos. Elencam-se, continuadamente, algumas das entidades não anotadas, assim como uma relação dos motivos pelos quais o *script* não funciona nestas ocasiões:

- Ausência de determinadas unidades lexicais no pacote lexical que se toma como ponto de partida: trata-se, fundamentalmente, de compostos que não foram incluídos no pacote lexical, embora o segundo dos seus elementos (a base) faça parte do pacote lexical. Um exemplo claro é o caso do lexema *Gruppe*, que é identificado como {animado, humano} e etiquetado corretamente. Não obstante, o recurso à expressão regular `.*gruppe` evidencia que existem outras entidades, como *Zielgruppe*, desprovidas de anotação automática.
- Aparição de particípios I para a realização do complemento sujeito: a execução da expressão regular `.*nden` permite-nos inferir que unidades lexicais como *Teilnehmenden* ou *Anwesenden* não são anotadas semanticamente por não estarem incluídas no pacote lexical {animado, humano}, embora outras palavras muito próximas quer do ponto de vista

formal quer do ponto de vista semântico, como *Teilnehmer*, já façam parte do pacote lexical.

- Presença de unidades lexicais difíceis de desambiguar: tal como defendem Apresjan (1974) e Renau (2021) trata-se de unidades lexicais que, do ponto de vista semântico ou lexical, podem ser consideradas como polissémicas, mesmo que os distintos significados de um significante estejam muito próximos entre si e só surjam por extensão semântica através de fenómenos como a metáfora ou a metonímia. Destaca-se a concordância *Frage der Zeitschrift*, onde o lexema *Zeitschrift* pode ser interpretado de duas formas diferentes:
    - '[[instituição]] encarregue pela publicação de uma revista'
    - '[[texto]] que resulta do processo de publicação realizado por uma organização ou instituição e que contém informação sobre diferentes tópicos'

  Considerando estas duas possibilidades de significado e de interpretação, deriva que a frase nominal *Frage der Zeitschrift* pode ser interpretada como 'pergunta da revista [instituição]' ou 'pergunta incluída na revista [texto]'. São casos muito concretos para os quais nem sempre é possível delimitar se uma anotação semântica com o *tag* {animado, humano} é adequada.

- Existência de erros tipográficos ou falhas na fase do pré-processamento: este elemento diz respeito, sobretudo, ao ruído computacional que aparece nas concordâncias escolhidas. Assim, por exemplo, casos como `<kwic>Frage der &quot;&quot;&quot; </kwic>` não podem ser anotados pela presença de elementos que dificultam a sua interpretação e leitura por parte da máquina. O mesmo acontece com as várias formas para grafar o feminino plural, pois ocorrências como `<kwic>Fragen der JournalistInnen</kwic>` ou `<kwic>Fragen der Teilnehmer_innen</kwic>` não são etiquetadas automaticamente.

Não devemos esquecer, todavia, outros elementos que nos permitem avaliar o correto funcionamento do *script* consoante os objetivos perseguidos. Trata-se, nomeadamente, de duas situações fundamentais:

- Correspondência de unidades lexicais não anotadas com outro padrão valencial: a estrutura em genitivo pode também desempenhar a função de complemento objeto e, com esta realização costuma ativar-se a segunda aceção *supra* citada ('problema, assunto'). Portanto, a extração de todos os elementos com a estrutura *Frage der/des*

[[humano]] faz com que fique uma lista ainda considerável com outras concordâncias com a mesma estrutura formal, cujos traços semânticos poderiam ser classificados, em geral, como [imaterial]. Estruturas como *Frage des Lebens*, *Frage des Geldes* ou *Frage der Erziehung* ocorrem muito frequentemente no *corpus* de trabalho.

- Correspondência das entidades não anotadas com realizações de adjuntos ou circunstantes: abordam-se agora situações em que o genitivo corresponde com a realização formal de circunstantes. A estrutura *Frage der Zeit* representa uma percentagem de 8,34% das ocorrências totais, facto que aponta para a sua elevada frequência. Isto demonstra mais uma vez que a frequência não pode ser considerada como o único critério para a análise de padrões lexicais.

Estas duas últimas situações mostram, em consequência, que o funcionamento do *script* é adequado, pois permite extrair entidades anotadas com o *tag* selecionado, de modo a facilitar a manipulação e análise das restantes, que poderiam ser anotadas posteriormente com outro pacote lexical. Aliás, no caso do predicado nominal *Frage*, consegue-se, pelo menos parcialmente, uma desambiguação de significados, pois como já se afirmou, a estrutura em genitivo com o traço semântico [humano] costuma associar-se à primeira aceção (cf. Sommerfeldt & Schreiber, 1983b). Para além disso, o pacote lexical de trabalho pode ser retroalimentado com a inclusão das entidades que não foram anotadas e se considerem adequadas.

# 7

## Conclusão

A presente dissertação constitui uma primeira abordagem da questão da anotação semântica (semi)automática em *corpora* linguísticos, nomeadamente para a língua alemã. Neste capítulo, apresentar-se-á uma reflexão sobre o trabalho conduzido, assim como sobre a fiabilidade da metodologia escolhida. Em primeiro lugar, sintetizar-se-á o quadro em que se deve situar esta investigação.

Como já referido, o objetivo principal desta pesquisa consistia em avaliar a efetividade de anotar *corpora* linguísticos do ponto de vista semântico-ontológico ou semântico-categorial, já que na literatura científica se encontram referências à dificuldade de extrair alguns dados linguísticos devido à falta de anotação semântica (*vd*. Domínguez, 2014; Domínguez *et al*., 2018; López, 2020). Para tal propósito, delimitou-se a frase nominal como sendo a unidade linguística mínima com significado pelo facto de as unidades lexicais aparecerem integradas num contexto (*vd*. Gross, 2013).

Neste sentido, escolheu-se o campo lexical da comunicação (cf. Hernández, 1993) como âmbito de aplicação e a partir da sua descrição do ponto de vista teórico (*vd*. Geckeler, 1971; Coseriu, 1977), pôde-se realizar uma seleção de substantivos que pertencessem a este campo para a sua análise posterior. A gramática e lexicografia de valências (*vd*. Domínguez, 2011; PORTLEX) desempenharam, aliás, um papel fundamental na delimitação deste estudo, pois partiu-se da estrutura argumental para a proposta de sistema automático de anotação semântica.

Assim, a combinação de diferentes teorias para a análise do léxico e da coocorrência semântica (*vd*. Fillmore, 1982; Hanks, 2004, 2013) com os resultados obtidos no âmbito de projetos como MultiGenera e MultiComb para a geração automática de estruturas e descrição de traços semânticos, foi essencial para o desenvolvimento deste trabalho. De facto, a ontologia lexical (Domínguez *et al*., 2021) converteu-se no ponto de partida para a delimitação do pacote léxico com que se operou nesta dissertação.

Igualmente, no concernente ao enquadramento metodológico, juntaram-se aplicações provenientes de diferentes áreas, nomeadamente da linguística de *corpus*, da linguística computacional e da inteligência artificial. Destarte, propôs-se um *script* desenhado *ad hoc* para anotar semanticamente em linguagem de marcação XML o complemento sujeito na sua realização como frase genitiva para três predicados nominais em alemão (*Bericht*, *Diskussion*, *Frage*). Para tal fim, tornou-se necessário recorrer

à anotação formal através de consultas CQL subjacente aos *corpora* integrados em Sketch Engine e que nos permite peneirar dados linguísticos estabelecendo determinadas estruturas morfossintáticas.

Um cruzamento dos dados extraídos de corpora com o script desenhado nesta pesquisa permitiu-nos retirar conclusões estatísticas sobre a viabilidade de sistematizar a etiquetagem semântica para *corpora* linguísticos. Estes resultados foram apresentados de forma pormenorizada em § 6 e em seguida, resumir-se-á a informação mais relevante para a conclusão da presente pesquisa.

Se se tomarem em consideração parâmetros estatísticos, deduz-se que o *script* criado permitiu anotar, de forma automática, uma média ($\bar{x}$) de 50,33% das ocorrências de unidades lexicais que podem ser classificadas com a etiqueta {animado, humano}. Não obstante, a média seria maior se não se incluíssem na contagem os resultados obtidos para as estruturas argumentais com o substantivo *Bericht* como núcleo.

No caso concreto do nome *Bericht*, a aparição de siglas (9%) e de unidades lexicais que não foram integradas no pacote lexical de trabalho, por se referirem amiúde a instituições ou organizações específicas, fez com que a percentagem de concordâncias não anotadas automaticamente fosse maior, cerca de 62%. Porém, o recurso a técnicas semiautomáticas, como a utilização de expressões regulares, permitiu-nos deduzir que algumas das unidades lexicais não anotadas poderiam ser empregues posteriormente para retroalimentar e alargar o pacote lexical já criado. Tal é o caso de siglas como *HNA* ou *NDR*, para além de muitos compostos em que o elemento base é a palavra *Vorsitzende*, por exemplo.

Para além disso, a análise realizada nesta dissertação, permite-nos concluir, do ponto de vista linguístico, que a realização em genitivo para o complemento sujeito de *Bericht* acostuma contar com o traço [instituição] e esta realização formal para o actante agente é notoriamente mais frequente neste nome do que nos outros selecionados. Em suma, para *Bericht* contaram-se 4213 ocorrências com o traço semântico-ontológico {animado, humano}, enquanto o número calculado para *Diskussion* era de 1043 concordâncias e para *Frage*, de 774.

Portanto, se se excluísse o caso de *Bericht*, a média ($\bar{x}$) de fiabilidade do *script* desenhado ascenderia até uma percentagem aproximada de 79%, o que evidencia um correto funcionamento consoante os parâmetros pré-estabelecidos. Deve-se esclarecer, todavia, que o pacote lexical poderia ser afinado de modo a incluir um maior número de unidades lexicais, assim como de modo a aumentar a granularidade da ontologia lexical já existente.

Deste modo, a partir do exposto ao longo deste trabalho, infere-se que se consegue automatizar, através da criação de um *script* informático, um sistema para anotarmos semanticamente

unidades lexicais em corpora linguísticos. Sendo afirmativa a resposta à primeira questão de investigação (*vd.* § 5.1.), cabe salientar que a elevada percentagem de exatidão torna razoável continuar com esta linha de pesquisa. Não se deve esquecer, contudo, que ocorreram ao longo da investigação uma relação de dificuldades:

- É necessário alargar e afinar o pacote lexical de trabalho, pela falta de representatividade de algumas unidades lexicais, nomeadamente siglas e compostos.
- É preciso desenvolver um sistema de desambiguação para unidades que se caracterizam pela sua polissemia regular (*vd.* Renau, 2021).
- É imprescindível estabelecer uma fase de pré-processamento para apagar erros tipográficos que existem nos *corpora* com que se trabalha.

Assim, também se corrobora que este sistema de anotação semântica abre novos caminhos para a desambiguação sintática (*vd.* Móia, 2016) e até semântica de algumas unidades lexicais que ativam um significado específico quando combinadas com traços semânticos específicos. Isto pôde-se verificar especificamente no caso de *Frage*, pois o esquema *Frage der/des* [[humano]] refere-se habitualmente à aceção do lema significando 'pergunta', enquanto a combinação com outro traço semântico ativa frequentemente o sentido 'assunto, problema'.

Deve-se ainda sublinhar que as consultas CQL morfossintáticas, embora ajudem a peneirar informação morfossintática, nem sempre são suficientes para a recolha de dados linguísticos específicos, como a descrição de estruturas argumentais. Isto também acontece com critérios quantitativos como a frequência, já que, por vezes, a informação que um utente necessita acerca de um sistema linguístico não corresponde com a realização mais frequente (*vd.* Domínguez *et al.*, 2021).

Em termos gerais, pode-se concluir que este trabalho alcançou os objetivos estabelecidos em § 2, pois ao longo da dissertação o foco centrou-se em algumas carências atuais da anotação existente em *corpora*, fundamentalmente as relativas às consultas CQL morfossintáticas ou formais, que nem sempre providenciam a informação linguística desejada por não contarem com filtro semântico. Nesta linha, desenvolveu-se um protótipo de *script* que permite etiquetar semanticamente de forma automática o complemento sujeito em genitivo. O recurso assíduo a ferramentas de PLN conduz à consolidação de novas ferramentas na linguística aplicada e na lexicografia. Abrem-se ainda, neste sentido, novos caminhos para a criação de recursos que permitam a desambiguação semântica de estruturas linguísticas e unidades lexicais.

Para consequentes estudos, seria desejável alargar o âmbito de aplicação, escolhendo nomes de outros campos lexicais e criando novas listas de vocabulário (pacotes lexicais), para abrir a porta à anotação de outros complementos no nível da frase nominal. Desde que se demonstra que o *script* desenhado funciona, a manipulação posterior é mais fácil e pode-se realizar uma sistematização e operacionalização do recurso criado. Assim, ambiciona-se, para projetos futuros, continuar a trabalhar em outras possibilidades de desambiguação da polissemia partindo da semântica combinatória (veja-se o exemplo: *die Frage der Teilnehmenden* vs. *die Frage der Gerechtigkeit*). Desta forma, conseguir-se-ia preencher duas lacunas importantes: a falta de anotação semântica sistematizada em *corpora* linguísticos e a existência de poucos recursos que permitam uma desambiguação de estruturas automaticamente.

## Referências bibliográficas

### a. Literatura científica


Ágel, V. (2000). *Valenztheorie*. Narr.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics, 12(142)*, 5-32.

Arias A., I. (2020). *Über nicht valenzbedingte Modifikationen bei nominalen Argumentstrukturen des semantischen Feldes der Kommunikation* (Dissertação de Licenciatura, Universidade de Santiago de Compostela). Minerva, Repositório Institucional. http://hdl.handle.net/10347/24096

Baker, C. (2012). FrameNet, current collaborations and future work. *Lang Resources & Evaluation, 46*, 269-286. doi: 10.1007/s10579-012-9191-2.

Bassola, P. (2003). *Deutsch-ungarisches Wörterbuch zur Substantivvalenz*. Szeged.

Bassola, P., Kubczak, J. & László, S. (2004). Zweisprachige Substantivvalenz in Theorie und Praxis. In S. Stănescu (ed.), *Die Valenztheorie. Bestandsaufnahme und Perspektiven. Beiträge der Tagung in Hermannstadt/Sibiu vom 19. bis 24. Februar 2002, Frankfurt/M.* (pp. 175-184).

Bernardos, M. S. (2007). ¿Qué es la generación de lenguaje natural? Una visión general sobre el proceso de generación. Inteligencia Artificial. *Revista Iberoamericana de Inteligencia Artificial, 11(4*), 105-128.

Brinkmann, H. (1971). *Die deutsche Sprache. Gestalt und Leistung*. Schwann.

Calderón C., M. (1994). *Sobre la elaboración de diccionarios monolingües de producción. Las definiciones, los ejemplos y las colocaciones léxicas*. Universidad de Granada.

Coseriu, E. (1977). *Principios de semántica estructural*. Editorial Gredos.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Curcio, M. L. (1999). *Kontrastives Valenzwörterbuch der gesprochenen Italienisch-Deutsch*. Institut für Deutsche Sprache.

Domínguez V., M. J. (2011). *Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens*. Iudicum.

Domínguez V., M. J. (2014a). Das Verb und seine Mitspieler: die häufig vergessene semantische Ebene. *AION: annali, sezione germanica, nuova serie, XXIV 1(2)*, 59-78.

Domínguez V., M. J. (2014b). Nomenergänzungen aus grammatischer Sicht. Forschungsstand und Bestandsaufnahme. *Neuphilologische Mitteilungen, 115(1)*, 3-32.

Domínguez V., M. J. (2021). Zur Darstellung eines mehrstufigen Prototypbegriffs in der multilingualen automatischen Sprachgenerierung: vom Korpus über *word embeddings* bis hin zum automatischen Wörterbuch. *Lexikos 31 (AFRILEX-reeks/series 31)*, 20-50

Domínguez V., M. J., Valcárcel R., C. & Lindemann, D. (2018). Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantical Knowledge from Corpora (MultiGenera). Em J. Čibej,

V. Gorjanc, I. Kosem & S. Krek (eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 847-854). Ljubljana University Press.

Domínguez V., M. J., Solla P., M. A. & Varcálcel R., C. (2019). Resources interoperability: exploiting lexicographic data to automatically generate dictionary exemples. Em *Proceedings of the VI. eLex conference. Electronic Lexicography in the 21st century* (pp. 51-71). Smart Lexicography.

Domínguez V., M. J. & Valcárcel R., C. (2019). PORTLEX as a multilingual nad cross-lingua online dictionary. In M. J. Domínguez, M. Mirazo & C. Varcárcel (eds.), *Studies on Multilingual Lexicography* (pp. 135-158). De Gruyter.

Domínguez V., M. J. (2020). Aplicación de WordNet e de word embeddings no desenvolvemento de prototipos para a xeración automática da lingua. *Linguamática, 12(2)*, 71-80.

Domínguez V., M. J., Bardanca O., D. & Simões, A. (2021). Automatic Lexicographic Content Creation for Lexicographers. *Proceedings of eLex 2021*, 269-287.

Dušek, O. (2013). *Zum Vergleich der tschechischen und deutschen Valenzwörterbücher* (Dissertação de Mestrado). Univerzita Karlova v Praze. https://ufal.mff.cuni.cz/~odusek/theses/ma_thesis.pdf

Eisenberg, P. (2001). *Grundriss der deutschen Grammatik*. Metzler.

Engel, U. (1996). Semantische Relatoren. Ein Entwurf für künftige Valenzwörterbücher. Em N. Weber (ed.), *Semantik, Lexikographie und Computeranwendungen* (pp. 223-236). Niemeyer.

Engel, U. (2004). *Deutsche Grammatik – Neubearbeitung*. Iudicum.

Engel, U. & Savin, E. (1983). *Valenzlexikon Deutsch-Rumänisch*. Julius Groos Verlag.

Engelberg, S. (2019). Argumentstrukturmuster. Ein elektronisches Handbuch zu verbalen Argumentstrukturen im Deutschen. Em D. Czicza, V. Dekalo & G. Diewald (eds.), *Konstruktionsgrammatik IV. Varianz in der konstruktionalen Schematizität* (pp. 13-38). Staffenburg.

Eroms, H. W. (2000). *Syntax der deutschen Sprache*. De Gruyter.

Fillmore, C. H. (1977). Scenes-and-frames semantics. Em Zampolli, A. (ed.), *Linguistic Structures Processing*, (pp. 55-79). North Holland Publishing.

Fillmore, C. J. (1982). Frame semantics. *Linguistics in the Morning Calm*, 111-137.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica, 6(2)*, 222-254.

Fuertes-Olivera, P. A. & Bergenholtz, H. (2018). Dictionaries for text production. Em Fuertes-Olivera, P. A. (ed.), *The Routledge Handbook of Lexicography*, (pp. 267-283). Routledge.

Geckeler, H. (1971). *Strukturelle Semantik und Wortfeldtheorie*. Wilhelm Fink.

Golonka, J. (2002). *Ihre Meinung dazu* oder: *Wie denken Sie darüber*? Zur Vererbung verbaler Valenzmerkmale in Nominalphrasen des Deutschen und des Polnischen. *Arbeitspapiere und Materialien zur deutschen Sprache, 2(2)*.

González-Agirre, A. & Rigau, G. (2013). Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository. *Linguamática, 5(1)*, 13-28.

Gómez G., X. & Solla P., M. A. (2019). Construction of a WordNet-based multilingual lexical ontology for Galician. In M. J. Domínguez, M. Mirazo & C. Varcárcel (eds.), *Studies on Multilingual Lexicography* (pp. 179-196). De Gruyter.

Gross, G. (2013). *Manual de análisis linguístico*. Editorial UOC.

Hanks, P. (2004). Corpus Pattern Analysis. *EURALEX 2004 Proceedings*, 87-97.

Hanks, P. (2013). *Lexical Analysis. Norms and Exploitations*. MIT Press.

Helbig, G. (1992). *Probleme der Valenz und Kasustheorie*. Niemeyer.

Helbig, G. & Schenkel, W. (1969). *Wörterbuch zur Valenz und Distribution deutscher Verben*. De Gruyter.

Hernández E., J. (1993). *Verba dicendi. Kontrastive Untersuchungen Deutsch-Spanisch*. Peter Lang.

Hölzner, M. (2007). *Substantivvalenz. Korpusgestützte Untersuchungen zu Argumentrealisierungen deutscher Substantive*. De Gruyter.

Iriarte, Á. (2001). *A Unidade Lexicográfica. Palavras, Colocações, Frasemas, Pragmatemas*. Braga: Centro de Estudos Humanísticos da Universidade do Minho.

Iriarte S., Á. (2004). Dicionários codificadores. Em C. M. de Sousa & R. Patrício (eds.), *Largo mundo alumiado: estudos em homenagem a Vítor Aguiar e Silva*, vol. 1, (pp. 81-98). Centro de Estudos Humanísticos da Universidade do Minho (CEHUM). http://hdl.handle.net/1822/3318

Jakobson, R. (1971). *Selected Writings: Word and Language*. Mouton.

Jakubíček, M., Kilgarriff, A., McCarthy, D. & Rychly, P. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. *PACLIC*, 741-747.

Kleiber, G. (1998). *Prototypensemantik. Eine Einführung*. Gunter Narr.

Kubczak, J. & Constantino, S. (1998): Exemplarische Untersuchungen für ein syntagmatisches Wörterbuch Deutsch – Französisch/Französisch – Deutsch. Em D. Bresson & J. Kubczak (eds.): *Abstrakte Nomina. Vorarbeiten zu ihrer Erarbeitung und Erfassung in einem zweisprachigen Wörterbuch*, (pp. 11-119). Gunter Narr.

Kubczak, J. & Schumacher, H. (1998). Verbvalenz – Nominalvalenz. Em D. Bresson & J. Kubczak (eds.), *Abstrakte Nomina. Vorarbeiten zu ihrer Erfassung in einem zweisprachigen syntagmatischen Wörterbuch* (pp. 273-286). Gunter Narr.

López I., N. (2020). Analysing nominal phrase context for the automatic extraction of linguistic and lexicographic data (Dissertação de Mestrado, Universidade do Minho). RepositoriUM. http://hdl.handle.net/1822/68562

Lutzeier, P. R. (1981). *Wort und Feld: wortsemantische Fragestellungen mit besonderer Berücksichtigung des Wortfeldbegriffes*. Niemeyer.

Lyons, J. (1977). *Semântica* (Volume 1). Editorial Presença.

Marek, T. (2009). Integration of light-weight semantics into a syntax query formalism. An extension of the tiger query language (Dissertação de Mestrado, Universidade de Saarland). https://www.coli.uni-saarland.de/projects/salsa/papers/torstenMSc.pdf

Matsekh-Ukrayinskyy, L. (2015). *Adjektivvalenz und präpositionale Komplemente. Eine framebasierte Untersuchung zu Syntax und Semantik der präpositionalen Komplemente bei Adjektiven*. Peter Lang.

Mestrado Europeu em Lexicografia (2018). *Teses Villa Vigoni*. https://www.emlex.phil.fau.de/ueberuns/publikationen/andere-publikationen/

Ministério de Educação e Ciência (2008). *DT: Dicionário Terminológico para consulta em linha*. https://dt.dge.mec.pt/

Mirazo, M. (2016). El e-diccionario multilingüe de la valencia del sustantivo PORTLEX. Algunas dificultades técnicas y metodológicas en la elaboración de su diseño y estructura. Em A. Castell (ed.), *Sintaxis y diccionarios: la complementación en alemán y español* (pp. 89-116). Peter Lang.

Móia, T. (2016). Semântica e Pragmática. Em A. M. Martins & E. Carrillo (eds.), *Manual de Linguística Portuguesa, Manuals of Romance Linguistics 16*. De Gruyter.

Porto Editora (2009). *Dicionário da Língua Portuguesa*. Dicionários Editora.

Renau, I. & Nazar, R. (2016a). Automatic Extraction of Lexical Patterns from Corpora. *Proceedings of the XVII EURALEX International Congress*, 823-830.

Renau, I. & Nazar, R. (2016b). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 1485-1492.

Renau, I. (2021). Algunos datos lexicográficos y de corpus para la representación de la polisemia regular en los diccionarios. Em San Martín N., A., Rojas G., D. & Chávez F., S. (eds.), *Estudios en homenaje a Alfredo Matus Olivier. Volumen II: Anejo N°3 Boletín de Filología*, (pp. 905-926). Facultad de Filosofía y Humanidades, Universidad de Chile.

Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F. & Roventini, A. (1998). The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. *Computers and the Humanities, 32*, 117-152.

Saussure, F. (1976). *Curso de Linguística Geral* (2ª edição). Dom Quixote.

Schierholz, S. J. (2001), *Präpositionalattribute. Syntaktische und semantische Analysen*. De Gruyter.

Schierholz, S. J. (2008). Corpusbasierte Operationalisierungsstrategien zur Bestimmung von Valenzpartnern. Em Donhauser, K. (ed.), Em*pirische Grundlagen moderner Grammatikforschung (Band 4)*, (pp. 37-48). Peter Lang.

Schierholz, S. J. (2013). Ein Online-Wörterbuch zur Valenz der Substantive. Em V. Jesenšek (ed.), *Lexicography. Print and Digital, Specialised Dictionaries, Databases*, (pp. 95-112).

Schumacher, H. (1986). *Verben in Feldern*. De Gruyter.

Sommerfeldt, K. E. & Schreiber, H. (1983a). *Wörterbuch zur Valenz und Distribution deutscher Adjektive*. De Gruyter.

Sommerfeldt, K. E. & Schreiber, H. (1983b). *Wörterbuch zur Valenz und Distribution der Substantive*. Niemeyer.

Stanescu, S. (2008). Lehren und Lernen von Verben, Adjektiven und Substantiven... Ein nimmerendender Diskussionsstoff. Em Stanescu, S. & Engel, U. (eds.), *Sprachvergleich – Kulturvergleich. Quo vadis, KGdr?* (pp. 28-42). Iudicum.

Storrer, A. (2003). Ergänzungen und Angaben. Em V. Ágel, L. M. Eichinger, H. W. Eroms, P. Hellwig, H. J. Heringer & H. Lobin (eds.), *Dependenz und Valenz. Ein internationales Handbuch zur zeitgenössischen Forschung, 1. Halbband* (pp. 764-780). De Gruyter.

Teubert, W. (1979a). *Valenz des Substantivs. Attributive Ergänzungen und Angaben*. Schwann.

Teubert, W. (1979b). Ergänzungen und Angaben beim Substantiv. *Mitteilungen des IdS, 5*, 17-26.

Teubert, W. (2001). Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics, 6(1)*, 125-153.

Tesnière, L. (1959). *Eléments de syntaxe structural*. Klincksieck.

Tesnière, L. (2015). *Elements of structural syntax*. John Benjamins.

Valcárcel R., C. (2016). Las construcciones $N_1N_2$ como realizaciones actanciales del sustantivo en francés y su tratamiento en el diccionario multilingüe PORTLEX. Em Domínguez V., M. J. & Kutscher, S. (eds.), *Interacción entre gramática, didáctica y lexicografía: Estudios contrastivos y multicontrastivos*, (pp. 193-208). De Gruyter.

Valcárcel R., C. & Domínguez V., M. J. (2016). *Teste 'muerte': falantes a avaliar a aceitabilidade de frases nominais geradas artificialmente*. Blogue de Carlos Valcárcel Riveiro. https://carlosvalcarcel.net/2016/11/30/teste-muerte-falantes-a-avaliar-a-aceitabilidade-de-frases-nominais-geradas-artificialmente/

Wiegand, H. E. (1998). *Wörterbuchforschung. 1. Teilband*. De Gruyter.

Wierzbicka, A. (1996). *Semantics. Primes and Universals*. Oxford University Press.

Zifonun, G. (2003). Grundlagen der Valenz. Em V. Ágel, L. M. Eichinger, H. W. Eroms, P. Hellwig, H. J. Heringer & H. Lobin (eds.), *Dependenz und Valenz. Ein internationales Handbuch zur zeitgenössischen Forschung, 1. Halbband* (pp. 352-377). De Gruyter.

Zgusta, L. (1971). *Manual of Lexicography*. Mouton.

### b. Recursos em linha

CombiContext = Domínguez V., M. J., Valcárcel R., C., Bardanca O., D., Calañas C., J. A., Catalá T., N., Martín G., R., Mirazo B., M., Sanmarco, B., M. T. & Pino S., L. (2021). CombiContext. Prototipo online para la generación automática de contextos frasales y oracionales de la frase nominal em alemán, español y francés. Universidade de Santiago de Compostela. http://portlex.usc.gal/combinatoria/verbal [Consultado: 15/05/2022].

Combinatoria = Domínguez V., M. J., Valcárcel R., C., Bardanca O., D., Calañas C., J. A., Catalá T., N., López I., N., Martín G., R., Mirazo B., M., Sanmarco, B., M. T. & Pino S., L. (2020). Combinatoria. Prototipo online para la generación biargumental de la frase nominal en alemán, español y francés. Universidade de Santiago de Compostela. http://portlex.usc.gal/combinatoria/ [Consultado: 15/05/2022].

DUDEN = Duden (n. d.). Duden Onlinewörterbuch. https://www.duden.de/woerterbuch [Consultado: 01/06/2022].

DWDS = Berlin-branderburgische Akademie der Wissenschaften (n. d.). Digitales Wörterbuch der deutschen Sprache (DWDS). https://www.dwds.de/ [Consultado: 01/06/2022].

FrameNet = International Computer Science Institute in Berkeley (n. d.). FrameNet. https://framenet.icsi.berkeley.edu/fndrupal/ [Consultado: 30/04/2022].

Ontologia lexical = Domínguez V., M. J., Valcárcel R., C., Bardanca O., D. (2021). Ontología léxica. http://portlex.usc.gal/ontologia/. [Consultado: 02/06/2022].

OWID = Leibniz-Institut für Deutsche Sprache (2008). Online-Wortschatz-Informatioonssystem Deutsch (OWID). https://www.owid.de/index.jsp. [Consultado: 31/05/2022].

PDEV = Hanks, P. (2022) Pattern Dictionary of English Verbs. https://pdev.org.uk/ [Consultado: 30/04/2022].

PORTLEX = Domínguez V., M. J., Valcárcel R., C., Mirazo B., M., Sanmarco B., M. T., Simões, A. & Vale, M. J. (2018). Portlex. Diccionario multilingue de la valencia del nombre. Universidade de Santiago de Compostela. http://portlex.usc.gal/portlex/ [Consultado: 20/05/2022].

Sketch Engine = Lexical Computing CZ S.R.O. (n. d.). Sketch Engine. https://www.sketchengine.eu/ [Consultado: 15/05/2022].

Xera = Domínguez V., M. J., Valcárcel R., C., Bardanca O., D., Calañas C., J. A., Catalá T., N., López I., N., Martín G., R., Mirazo B., M., Sanmarco, B., M. T. & Pino S., L. (2020). Xera. Prototipo online para la generación monoargumental de la frase nominal en alemán, español y francés. Universidade de Santiago de Compostela. http://portlex.usc.gal/combinatoria/usuario [Consultado: 15/05/2022].

WordNet = Princeton University (n. d.). WordNet. A Lexical Database for English. https://wordnet.princeton.edu/. [Consultado: 10/05/2022].

## Anexos

Os ficheiros XML anotados semanticamente automaticamente com o *script* desenhado no âmbito desta dissertação para as estruturas argumentais dos nomes *Bericht*, *Diskussion* e *Frage* podem ser encontrados no seguinte repositório: https://gitlab.com/ivanariasarias/dissertacao-mestrado-emlex