Universidade do Minho
Escola de Letras, Artes e Ciências Humanas

João Vieira Antunes

Processamento linguístico de narrativas produzidas por crianças lusodescendentes e proposta de interface de pesquisa

abril de 2024

Processamento linguístico de narrativas produzidas por crianças lusodescendentes e proposta de interface de pesquisa
João Vieira Antunes

UMinho | 2024

Universidade do Minho
Escola de Letras, Artes e Ciências Humanas

João Vieira Antunes

Processamento linguístico de narrativas produzidas por crianças lusodescendentes e proposta de interface de pesquisa

Dissertação de mestrado
Mestrado em Estudos Luso-Alemães

Mestrado de grau duplo, coordenado pela Universidade do Minho e pela Goethe Universität Frankfurt am Main

Trabalho realizado sob a orientação da
Professora Doutora Idalete Maria da Silva Dias
da
Professora Doutora Cristina Maria Moreira Flores
e da
Professora Doutora Esther Rinke-Scholl

abril de 2024

AGRADECIMENTOS

A realização desta dissertação representa o culminar de vários meses de trabalho intenso, esforço e dedicação, que não seria possível sem a ajuda e contributo de várias pessoas. Nesse sentido, apraz-me começar por agradecer à minha família, pois sem eles eu não estaria onde estou hoje. Um agradecimento especial à minha mãe Elvira, pela mulher guerreira que é e por todos os ensinamentos que me transmitiu ao longo da vida, ao meu pai Hélder, pelo incentivo e por sempre acreditar nas minhas capacidades, aos meus irmãos Hélio e Paulo e ao meu sobrinho Simão, pelo apoio, leveza e descontração com que preenchem o meu dia a dia. Foram eles o meu porto de abrigo nas fases menos boas deste processo e os primeiros a dar-me ânimo e força para que continuasse esta caminhada. Obrigado por tudo! E porque a família vai mais além dos laços consanguíneos, aproveito para deixar, também, umas palavras de agradecimento à minha família de acolhimento (e do coração) argentina, que há 8 anos entrou na minha vida e desde então tem permanecido e manifestado todo o seu carinho e apoio ao longo deste percurso – Jadra, Cacho, Tami, Lea, Juanse e Cachi.

Em segundo lugar, este trabalho não seria possível sem a orientação da Professora Doutora Idalete Dias, que rapidamente se prontificou a aceitar este desafio. A si, professora, o meu mais sincero agradecimento por ter sempre acreditado nas minhas capacidades, por todo o conhecimento transmitido e pelas orientações que foi dando, que em muito contribuíram para levar a bom porto este projeto. À Professora Doutora Cristina Flores por também ter acreditado na realização desta dissertação, pelas oportunidades que me brindou e que me possibilitaram a descoberta de novos horizontes e áreas de estudo, pelos conhecimentos transmitidos ao longo do mestrado, pelo apoio e incentivo e pela disponibilidade que sempre demonstrou, até mesmo fora de horas. À Professora Doutora Esther Rinke pela sua amabilidade, por ter sido um suporte fundamental durante o período de mobilidade em Frankfurt e por todos os conhecimentos transmitidos. A vós, caríssimas orientadoras, o meu mais sincero agradecimento.

Quero, também, agradecer a todos os professores do mestrado em Estudos Luso-Alemães que tive o prazer de conhecer, tanto na UMinho como na Goethe Universität, e que contribuíram para a minha formação e gosto pelas relações interculturais e linguísticas luso-alemãs. Agradeço, ainda, a ambas instituições de ensino – UMinho e Goethe Universität – pelo acolhimento e por possibilitarem esta experiência tão enriquecedora.

Um agradecimento aos meus colegas do mestrado pelo companheirismo e por todas as memórias e momentos vividos em Frankfurt e em Braga, em especial à Inês por todo o apoio e amizade.

Agradeço à minha querida amiga de infância, Ana, por ter sido, ao longo destes anos, a luz nos meus dias cinzentos, pelas conversas, por ouvir os meus desabafos, pelos risos, pelo apoio e incentivo e pela sua visão pragmática da vida, que em muitos momentos me ajudaram a enfrentar os meus problemas com mais nitidez. À Rita, pelas conversas, pelas caminhadas, por me ouvir e pelas palavras de apoio e incentivo. À Joana e à Maria pelos longos anos de amizade, pela preocupação e pelas palavras de carinho que sempre se fizeram sentir, apesar da distância. Às amizades que Aveiro me deu e que hoje são parte essencial da minha vida – Leitão, Mel, Ana Costeira, Sara, Ana Nogueira, Diana e Francisco –, porque foi aí que tudo começou.

A todos aqueles, que embora não estejam mencionados nesta folha de papel, mas que à sua maneira, sempre demonstraram carinho e preocupação e contribuíram para que, dia após dia, não me faltasse o ânimo para escrever esta dissertação.

Por fim, e como escreveu outrora a ilustre artista Violeta Parra, *¡Gracias a la vida que me ha dado tanto!*

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

Processamento linguístico de narrativas produzidas por crianças lusodescendentes e proposta de interface de pesquisa

Este projeto conjugará duas importantes áreas de estudo da linguística e das humanidades digitais, nomeadamente o bilinguismo e a análise e tratamento de corpora. Para tal, foram transcritas e analisadas 40 narrativas retiradas de um projeto de investigação, coordenado pela Professora Doutora Cristina Flores, investigadora no Centro de Estudos Humanísticos da Universidade do Minho. O estudo centrou-se em crianças lusodescendentes, que vivem na Suíça (cantão alemão), tendo estas sido submetidas a dois instrumentos de recolha de dados em português europeu (PE) e em alemão padrão (AP). Numa das fases do projeto, os participantes ouviram uma história e tiveram de a recontar nas duas línguas, produzindo as narrativas que foram analisadas nesta dissertação. Com efeito, o presente projeto de dissertação tem como objetivo contribuir para a criação de um corpus eletrónico de narrativas produzidas em PE e AP por falantes de herança e o seu processamento. Através de técnicas de processamento de linguagem natural, o corpus foi lematizado e etiquetado ao nível morfossintático [*part-of-speech tagging*] com recurso ao *Sketch Engine* – uma ferramenta de análise e gestão dos corpora. Estas duas camadas de informação linguística permitiram identificar e analisar padrões linguísticos específicos dos informantes em questão e ainda de preservar as próprias narrativas e disponibilizar um recurso para a comunidade científica. Na segunda parte do projeto, é apresentada uma prova-conceito de um protótipo de interface de pesquisa, que tem como objetivo o armazenamento destes dados linguísticos na sua versão anotada e não anotada e a disponibilização deste recurso para toda a comunidade linguística, contribuindo, assim, para a sustentabilidade e preservação deste tipo de recursos.

PALAVRAS-CHAVE: corpus de narrativas escritas, crianças bilingues, interface de pesquisa, português língua de herança, processamento de linguagem natural

ABSTRACT

Linguistic processing of narratives produced by children of Portuguese descent and proposal of a search interface

This project will combine two important areas of study within linguistics and digital humanities, specifically heritage bilingualism and the analysis and electronic processing of corpora. To this end, 40 narratives from a research project coordinated by Professor Cristina Flores, a researcher at the Centre for Humanistic Studies at the University of Minho, were taken, transcribed, and analysed. This research focused on children of Portuguese descent living in Switzerland (German canton), who were given two data collection instruments in European Portuguese (EP) and Standard German (AP). In one of the stages of the project, the participants had to listen to a story and had to retell it in both languages, producing the narratives analysed in this dissertation. That said, this thesis aims to contribute to the creation and processing of an electronic corpus of narratives produced in EP and AP by heritage speakers. Through natural language processing techniques, the corpus was lemmatized and tagged at the morphosyntactic level [*POS-Tagging*] using *Sketch Engine* – a tool for analysing and managing corpora. These two layers of linguistic data allowed to identify and analyse the specific linguistic patterns of the informants in question, to preserve the narratives *per se*, and to make available the resource to the scientific community. The second part of this dissertation presents a proof of concept of a search interface prototype, which intends to store this linguistic data in its annotated and unannotated versions and make the resource available to the entire linguistic community, thus contributing to the sustainability and preservation of this type of resource.

KEYWORDS: bilingual children, corpus of written narratives, natural language processing, search interface, Portuguese as heritage language

ZUSAMMENFASSUNG

Linguistische Verarbeitung von Erzählungen durch Kinder portugiesischer Abstammung und Vorschlag für eine Forschungsschnittstelle

In diesem Projekt werden zwei wichtige Forschungsbereiche der Linguistik und der digitalen Geisteswissenschaften kombiniert, nämlich die Zweisprachigkeit von Herkunftssprechern und die Analyse und Verarbeitung von Korpora. Zu diesem Zweck wurden 40 Erzählungen aus einem Forschungsprojekt, das von Frau Professor Cristina Flores, einer Forscherin am Zentrum für geisteswissenschaftliche Studien der Universität von Minho, koordiniert wurde, gesammelt, transkribiert und analysiert. Im Mittelpunkt dieses Forschungsprojekts standen Kinder portugiesischer Abstammung, die in der Schweiz (deutscher Kanton) lebten und denen zwei Instrumente zur Datenerhebung in europäischem Portugiesisch (EP) und Hochdeutsch (AP) zur Verfügung gestellt wurden. In einer der Projektphasen mussten die Teilnehmer eine Geschichte anhören und in beiden Sprachen nacherzählen, wodurch die in dieser Arbeit analysierten Erzählungen entstanden. Ziel dieser Arbeit ist es, einen Beitrag zur Erstellung und Bearbeitung eines elektronischen Korpus von Erzählungen zu leisten, die von Herkunftssprechern in EP und AP produziert wurden. Mit Hilfe von Techniken zur Verarbeitung natürlicher Sprache wurde der Korpus lemmatisiert und auf morphosyntaktischer Ebene [*POS-Tagging*] mit Hilfe von *Sketch Engine* – einem Tool zur Analyse und Verwaltung von Korpora – getaggt. Diese beiden Ebenen linguistischer Daten ermöglichten es, die spezifischen sprachlichen Muster der betreffenden Informanten zu identifizieren und zu analysieren, die Erzählungen als solche zu erhalten und die Ressource der wissenschaftlichen Gemeinschaft zur Verfügung zu stellen. Der zweite Teil dieser Dissertation stellt einen *proof of concept* eines Prototyps für eine Suchschnittstelle vor, der diese linguistischen Daten in ihren annotierten und nicht-annotierten Versionen speichern und der gesamten linguistischen Gemeinschaft zur Verfügung stellen soll, um so zur Nachhaltigkeit und Erhaltung dieser Art von Ressourcen beizutragen.

SCHLÜSSELWÖRTER: Forschungsschnittstelle, Korpus geschriebener Erzählungen, linguistische Datenverarbeitung, Portugiesisch als Herkunftssprache, zweisprachige Kinder

ÍNDICE

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

ÍNDICE DE ILUSTRAÇÕES

LISTA DE SIGLAS E ABREVIATURAS

2L1 – Aquisição simultânea

ANC – American National Corpus

AP – Alemão padrão

BFLA – Bilingual First Language Acquisition

BNC – British National Corpus

BoE – Bank of English

BSLA – Bilingual Second Language Acquisition

CAL2 – Corpus de Aquisição de L2

CHAT – Codes for the Human Analysis of Transcripts

CHILDES – Child Language Data Exchange System

CLAN – Computerized Language Analysis

CLEG – Corpus of Learner German

CLUL – Centro de Linguística da Universidade de Lisboa

COBUILD – Collins Birmingham University International Language Database

CRPC – Corpus de Referência do Português Contemporâneo

IA – Inteligência Artificial

ICLE – International Corpus of Learner English

KWIC – Key Word In Context

L1 – Língua materna/primeira língua

L2 – Língua segunda

L3 – Língua terceira

LAD – Language Acquisition Device

LC – Linguística de Corpus

LE – Língua Estrangeira

LH – Língua de Herança

LLC – The London-Lund Corpus of Spoken English

PC – Período Crítico

PE – Português europeu

PLH – Português Língua de Herança

PLN – Processamento de Linguagem Natural

QuesFEB – Questionário Sociolinguístico Parental para Famílias Emigrantes Bilingues

SGML – Standard Generalized Markup Language

STTS – Stuttgart-Tübingen-Tagset

TEI – Text Encoding Initiative

TTR – Type-Token Ratio

XML – Extensible Markup Language

INTRODUÇÃO

O bilinguismo representa uma das mais ricas heranças culturais que tem marcado presença acentuada nas sociedades contemporâneas, dado que a estimativa apontada por Crystal (2003) é de que metade da população mundial seja bilingue e que dois terços das crianças no mundo cresçam em ambientes bilingues (p. 17). Proveniente dos fluxos migratórios, o bilinguismo de herança designa a aquisição da língua de origem da família por parte do falante (Almeida & Flores, 2017). Este fenómeno em família com *background* migratório representa, por sua vez, uma forma de preservação da identidade cultural e das suas origens. No plano académico, a aquisição de línguas em contexto bilingue ganhou destaque e tem vindo a possibilitar a recolha de dados, ao longo dos anos, que têm servido de objeto de estudo a várias investigações no âmbito da linguística (Lynch, 2017, p. 46). Por conseguinte, o objeto de estudo desta dissertação apresenta semelhantes características, visto que se trata de um conjunto de narrativas produzidas por crianças lusodescendentes residentes em cantões alemães da Suíça. Assim, a interseção da linguística e das humanidades digitais torna possível explorar e integrar os domínios do bilinguismo de herança e da análise de *corpus* em dois mundos linguísticos distintos – o português europeu (PE) e o alemão padrão (AP).

Com efeito, os avanços tecnológicos, aliados ao interesse crescente em utilizar recursos eletrónicos para complementar abordagens mais tradicionais da análise da língua e da literatura, permitiram que áreas de conhecimento da linguística, como a Linguística de Corpus (LC), se desenvolvesse e adquirisse relevância dentro da comunidade científica. Desta forma, e tirando proveito dos progressos no domínio do processamento de linguagem natural (PLN), que estabeleceu uma afincada relação entre as máquinas e a linguagem humana, a LC está intrinsecamente ligada ao uso do computador, uma vez que os dados recolhidos são, atualmente, manuseados no plano digital. Assim, esta área da linguística ocupa-se da recolha, compilação e análise de *corpora* eletrónicos, que surgiu da necessidade sentida por linguistas de se apoiarem em evidências reais do uso da língua, com o objetivo de explorar padrões recorrentes da linguagem e tecer teorias sobre o funcionamento linguístico.

A presente dissertação de mestrado baseou-se na recolha e compilação de um corpus linguístico, constituído por 40 narrativas escritas, onde os informantes foram incumbidos de recontar uma história em PE e em AP, dando origem à criação de um corpus de aprendizagem bilingue. Depois

de recolhidas, as narrativas foram transcritas, transitando, assim, do formato físico (em papel) para o digital e posteriormente processadas, utilizando técnicas avançadas de PLN. Conservando a sua forma original, o corpus foi lematizado e etiquetado ao nível morfossintático, oferecendo um conjunto de dados linguísticos com várias camadas suscetíveis para análise. O seu processamento foi levado a cabo pela ferramenta *Sketch Engine*[1], que possibilitou a anotação dos textos, assegurando a preservação destes, enquanto constituem um recurso valioso para a comunidade científica.

Na segunda parte desta dissertação, é apresentada uma prova-conceito de um protótipo de interface de pesquisa. Esta interface está idealmente concebida para armazenar os dados linguísticos nas suas versões anotada e não anotada compilados para esta dissertação, com o objetivo de democratizar o acesso a este recurso linguístico no panorama académico. Desta forma, procede-se a uma sucinta abordagem sobre conceitos, objetivos e funcionalidades que integram a interface de pesquisa, que, numa segunda instância, contribuirá para a sustentabilidade do corpus, assegurando a sua preservação e facilitando uma utilização mais ampla, sublinhando assim a profunda importância deste e outros recursos linguísticos.

O trabalho encontra-se estruturado em 6 grandes capítulos, que, por sua vez, se subdividem em subcapítulos. No capítulo 1 é feita uma abordagem sobre o processo de aquisição da linguagem, onde são apresentados alguns períodos de investigação que contribuíram para o desenvolvimento das teorias mencionadas. Segue-se uma reflexão sobre a aquisição bilingue, que pretende explicar alguns conceitos subjacentes a essa área de investigação e fornecer um contexto detalhado sobre o perfil dos informantes que produziram as narrativas analisadas nesta dissertação.

O segundo capítulo versa sobre o corpus como recurso para análise linguística, onde são exploradas características, a sua usabilidade como ferramentas de pesquisa, as várias tipologias, com um enfoque particular nos corpora de aprendizagem. É feita, ainda, uma exposição sobre a recolha das narrativas que permitiram a compilação do corpus em estudo.

No capítulo 3 procede-se a uma análise sobre o tamanho do corpus e o processo de tokenização.

O capítulo 4 aborda as várias etapas de anotação e processamento do corpus, onde são exploradas as características da metainformação, lematização e etiquetação morfossintática.

[1] https://www.sketchengine.eu/

No capítulo 5 são apresentados os resultados obtidos do processamento eletrónico do corpus em estudo com base na ferramenta *Sketch Engine*. São, também, discutidos os resultados segundo o cruzamento de variáveis do perfil sociolinguístico dos informantes.

Por fim, no capítulo 6 procede-se à apresentação da prova-conceito do protótipo de interface de pesquisa.

CAPÍTULO I – ENQUADRAMENTO TEÓRICO

1.1. AQUISIÇÃO DA LINGUAGEM

O campo da aquisição da linguagem infantil é um campo que, ao longo dos anos, foi passando por várias mudanças relativamente aos métodos e às abordagens teóricas utilizadas para estudar o processo de aquisição da linguagem (Ingram, 1989, p. 7). No seu livro *First Language Acquisition: method, description and explanation*, Ingram (1989) considera importante, primeiramente, que se se debruce sobre os períodos da história dos estudos da linguagem infantil, que considera serem três: (1) *o período de estudos de diário* (de 1876 a 1926), (2) *o período de estudo de grandes amostras* (de 1926 a 1957) e (3) *o período de amostragem longitudinal de línguas* (de 1957 até ao presente). Cada um destes períodos caracteriza-se pelo desenvolvimento e aprofundamento de vários métodos e teorias, que nesse então, procuravam estudar o processo de aquisição da linguagem em crianças.

Nesta secção serão abordadas de forma sucinta as teorias e os autores mais importantes que contribuíram para o estudo da aquisição da linguagem infantil.

1.1.1. *Baby biographies*

O primeiro período (de 1876 a 1926) é, como o próprio nome indica, caracterizado por estudos de diário, onde o observador (normalmente um dos cuidadores da criança) anotava num diário os comportamentos mais notórios da criança durante o seu crescimento, com foco no processo de aquisição da linguagem (Ingram, 1898, p. 7). Estes comportamentos podiam ser palavras, atitudes, primeiros passos, entre outras observações (Lorandi et al., 2011, p. 145). É de realçar, que durante este período, a criança ganha um foco particular, uma vez que se torna no objeto de estudo central para determinar a forma como se desenvolve (Ingram, 1989, p. 7). Esta metodologia bastante popular neste período viria a ficar conhecida como "baby biographies", visto que fornecia uma base descritiva bastante rica para o campo da aquisição da linguagem infantil (Ingram, 1989, p. 8), assim como dados importantes sobre o desenvolvimento infantil analisado longitudinalmente, já que as crianças eram acompanhadas ao longo de vários anos (Ingram, 1989, p. 13). Não obstante, Ingram (1989) aponta que esta metodologia foi criticada por ser tendenciosa, pois sendo o progenitor o observador, embora conhecesse bem a criança, correr-se-ia o risco de que o próprio anotasse apenas os comportamentos que considerasse importantes

(p. 8). Ademais, muitos destes diários apresentavam alguns interregnos entre os registos, fazendo com que, muitas vezes, o estudo não fosse contínuo (Ingram, 1989, p. 8). Também o facto de que o número da amostra fosse consideravelmente reduzido durante este período e que estes estudos fossem, na sua maioria, meramente descritivos, carecendo de bases teóricas, impossibilitava a generalização de dados e conclusões sobre o processo de aquisição da linguagem em crianças (Ingram, 1989, p. 9). Ainda assim, os estudos de diários são considerados pelo autor uma base de dados importantíssima para o campo da aquisição da linguagem infantil pela vasta e rica literatura recolhida durante esse período.

1.1.2. Behaviorismo

O segundo período (de 1926 a 1957) foi essencialmente influenciado por uma abordagem behaviorista (Ingram, 1989, p. 12). Do inglês "behaviourist", o caráter desta teoria caracteriza-se por ser empirista, tendo como foco principal o estudo do comportamento [*behaviour*] (Ingram, 1989, p. 12). Em que é que este segundo período difere do primeiro acima descrito? A explicação é simples. Segundo Ingram (1989), o behaviourismo difere relativamente ao papel da criança na aprendizagem da língua e na medição do comportamento observável (p. 12). Isto é, segundo esta abordagem, a criança é passivamente controlada pelo seu ambiente, sendo os seus comportamentos influenciados pelo meio em que se insere e pela comunidade linguística envolvente, enquanto durante o primeiro período o papel ativo e espontâneo da criança era o foco principal. Um dos grandes estudiosos da teoria behaviourista foi B. F. Skinner, que influenciado pela filosofia da psicologia de John B. Watson, publicou, em 1957, o seu livro *Verbal Behavior*, onde viria a dar um grande contributo para esta abordagem. Skinner (1957) afirma sob esta perspetiva, que todo o comportamento – incluindo a linguagem – ocorre através de respostas dadas a uma série de estímulos (p. 31). Desde então e até aos anos 60, os estudos sobre a aquisição da linguagem eram fortemente determinados por uma abordagem behaviorista (Meisel, 2011, p. 3). Ao contrário do anterior, este foi um período em que os estudos realizados eram feitos de forma transversal, ou seja, estudos realizados a diferentes crianças de idades distintas (Ingram, 1989, p. 13). Também se diferenciou pela expressiva amostra analisada, já que o objetivo desta teoria não priorizava o acompanhamento da criança no seu processo de aquisição, mas sim a recolha do maior número de amostras e, consequentemente, traçar linhas concretas que

definissem um perfil padrão do comportamento das crianças durante o seu processo de aquisição da linguagem (Ingram, 1989, pp. 12-13).

1.1.3. Inatismo/nativismo

Fortemente apontada pelos linguistas modernos como uma teoria "pouco sofisticada" (Ingram, 1989, p. 16), o behaviorismo foi perdendo força, cedendo espaço para novos investigadores que, insatisfeitos com esta abordagem, continuaram o seu trabalho na procura de explicações para o processo de aquisição da linguagem. A famosa revisão de Chomsky (1959) do livro de Skinner (1957), *Verbal Behavior*, representa uma reviravolta nas ciências linguísticas durante este terceiro e último período identificado por Ingram (1989). Nesta nova abordagem apresentada por Chomsky, a aquisição da linguagem começa a ser encarada como uma atividade mental que ocorre no sistema cognitivo do indivíduo (Meisel, 2011, p. 3), contrariamente ao que Skinner argumentava. De acordo com a tese defendida pelo autor, o ser humano nasce com uma faculdade inata que lhe permite adquirir e produzir linguagem desde que esteja exposto a um *input* linguístico[2] (Chomsky, 1959, p. 42). Segundo a sua teoria inatista (Ingram, 1989, p. 25), o processo é biologicamente determinado: os circuitos neurais presentes no cérebro da espécie humana contêm informação linguística, sendo a predisposição natural da criança de aprender uma língua desencadeada pelo *input* linguístico a que é exposta desde a nascença (Chomsky, 1959, p. 42). Consequentemente, o cérebro da criança é capaz de interpretar o que ouve de acordo com as regras ou estruturas subjacentes que já contém (Chomsky, 1959, p. 42). Com isto, Chomsky (1959) não sugeriu que, por exemplo, uma criança nascida na China já nascesse a saber falar mandarim, mas sim que, uma vez que as línguas humanas partilham um certo conjunto de regras, é tarefa da criança estabelecer como a língua exprime essas regras subjacentes (p. 43). Com uma base teórica mais sólida, a *Language Acquisition Device* (LAD) proposta por Chomsky, como ficou conhecida na década de 60, conseguiu responder às inquietudes da comunidade linguística, tendo tido um enorme impacto na investigação da L1 e contribuindo para um número cada vez maior de publicações aplicando estas ideias ao estudo da aquisição da linguagem (Meisel, 2011, p. 3).

[2] Entenda-se, neste contexto, por *input* linguístico a língua a que uma criança é exposta durante o período de aquisição (Hoff-Ginsberg & Shatz, 1982). O *input* pode incluir tanto linguagem falada como escrita, bem como gestos e linguagem corporal. Este termo é frequentemente utilizado em contraste com *output*, que corresponde à linguagem que o aprendente produz.

1.1.4. Construtivismo

A investigação de Chomsky foi, de facto, uma alavanca que abriu horizontes nas áreas de estudo da linguística, psicolinguística e neurolinguística. Da teoria inatista, Ingram (1989) considera haver duas perspetivas que diferem ligeiramente uma da outra: a perspetiva maturacionista, visão que atribui a Chomsky e a perspetiva construtivista, visão viria a ser estudada e desenvolvida pelo psicólogo suíço Jean Piaget (p. 26). O trabalho desenvolvido por Piaget no âmbito da aquisição da linguagem destacou-se por várias razões, sendo as principais por consagrar um impressionante conjunto de factos sobre o desenvolvimento precoce da criança, por ir mais além da explicação dos factos e, mais importante, pelo seu próprio conteúdo (Ingram, 1989, p. 116). Piaget argumenta que o ser humano é capaz de criar conhecimento através da interação entre as suas experiências e ideias (Blumer, 1930, p. 151). Deste modo, durante o processo de aquisição da linguagem, a criança está no centro do processo de criação e aquisição do conhecimento. Na sua investigação, Piaget propõe quatro estágios de desenvolvimento da criança, sendo a aquisição da linguagem parte desse processo (Lorandi et al., 2011, p. 147). Acrescenta que, inerentes à aquisição da linguagem, estão os períodos sensório-motor (do nascimento ao segundo ano de vida) e o pré-operatório (dos dois aos sete anos de idade). A teoria de Piaget enfatiza, deste modo, a natureza inata dos três processos de assimilação, acomodação e de organização e como estes se baseiam em "reflexos inatos" para construir o conhecimento (Ingram, 1989, p. 117). Desta forma, para Piaget, o meio em que a criança vive contribui para que esta possa desenvolver e adquirir linguagem, assim como adquire qualquer conhecimento durante o seu processo de desenvolvimento.

1.1.5. Sociointeracionismo/Socioconstrutivismo

Também relacionada com o meio, mas com particular foco na interação social, está a teoria desenvolvida pelo psicólogo russo Lev S. Vygotsky, que estabeleceu que o desenvolvimento de um individuo resulta de um processo sócio-histórico (Lorandi et al., 2011, p. 148). Na sua teoria sociointeracionista, Vygotsky defende que a aquisição de conhecimentos se dá pela interação do sujeito com o meio. Desta forma, a questão central para o processo de aquisição da linguagem prende-se na interação e relação estabelecida entre o falante e o interlocutor, do que propriamente no resultado dessa interação (Lorandi et al., 2011, p. 148). Na mesma linha de pensamento, Jerome S. Bruner, que inspirado pela teoria de Vygotsky, sublinha a importância do contributo

linguístico que as crianças recebem dos seus cuidadores (Bruner, 1981, p. 162). Defende, ainda, que a língua existe para efeitos de comunicação e que esta é aprendida inevitavelmente em contexto de interação (Bruner, 1981, p. 159). O psicólogo estadunidense refere que o comportamento linguístico dos adultos, quando interagem com crianças, é essencialmente adaptado, de modo que a interação seja profícua para ambas as parte, e que desta forma, contribui para o processo de aquisição da linguagem da criança (Bruner, 1981, p. 175). A este modelo de interação, Bruner denomina de *Language Assistance System* em resposta ao LAD proposto por Chomsky nos anos 60 (Bruner, 1981, p. 169).

A estas teorias e metodologias somam-se muitas outras que, nos últimos cinquenta anos, têm vindo a investigar intensamente a aquisição da linguagem no âmbito da linguística, da psicolinguística e da psicologia do comportamento (Côrrea, 2018, p. 30). Com isto, não se pretende mostrar que apenas as teorias acima mencionadas respondem às principais questões sobre esta temática, nem que uma está mais acertada que a outra. Considera-se, pois, que o desenvolvimento e aprofundamento de todas as teorias foi essencial para que se continuasse a questionar sobre este processo transversal aos seres humanos, dando espaço para que cada autor, mediante a sua investigação, pudesse (ou tentasse) explicar como este processo de aquisição da linguagem ocorre. Em suma, de uma ou outra forma, estas teorias que tentam explicar como ocorre o processo de aquisição da linguagem complementam-se, existindo uma certa proximidade que, de forma tangível, as faça coincidir em alguns aspetos. Não obstante, esta secção serve apenas de contextualização à temática da aquisição bilingue, tema que será abordado doravante.

1.2. AQUISIÇÃO BILINGUE E BILINGUISMO

Apesar das várias teorias que procuram explicar o processo de aquisição da linguagem, é consensual dentro da comunidade linguística que a aquisição da linguagem se desenvolve naturalmente nas crianças e que estas adquirem igualmente as competências para a utilizar, sem que lhes sejam ensinadas (Meisel, 2011, p. 2). Também na mesma linha de pensamento, Bhatia (2006) refere que por questões inerentes à condição humana, uma criança (saudável) tem mais facilidade em adquirir qualquer linguagem humana do que um adulto, independentemente do seu género, raça, etnia ou nacionalidade (pp. 16-17). E as crianças bilingues? Como ocorre o processo de aquisição de duas (ou mais) línguas? Na presente secção será abordado como sucede a

aquisição de duas línguas por crianças e como os vários fatores, como a idade, o contexto de aquisição e a quantidade e qualidade de *input* linguístico recebidos podem ter influência neste processo.

Independentemente da motivação, do esforço, do domínio da sintaxe, entre outros aspetos, tem sido observado por vários linguistas que até mesmo os bilingues que detêm um excelente nível de proficiência nas suas primeira (L1) e segunda (L2) línguas, não são bilingues perfeitos (Bhatia, 2006, p. 17). Para explicar estas e outras dissemelhanças na aquisição da linguagem, Lenneberg propôs, em 1967, a teoria da "hipótese do período crítico" que relaciona a idade do falante com a facilidade de aquisição da linguagem (Bhatia, 2006, p. 17). No seu artigo *Biological Foundations of Language*, Lenneberg (1967) afirma a existência de um período de maturação ou período crítico (PC) entre os dois anos de idade e a puberdade (rondando os doze anos de idade), no qual é possível uma aquisição da linguagem pelo falante completa e quase sem esforço (p. 63). O autor justifica a sua afirmação por este ser o período em que ocorre a finalização da lateralização do cérebro (Lenneberg, 1967, p. 65). Também Meisel (2011) afirma que as crianças adquirem naturalmente a capacidade de falar sem grande esforço aparente e sem que lhes seja ensinada a língua, ao contrário dos adolescentes e (jovem) adultos que, quando aprendem uma língua estrangeira em contexto de sala de aula, por exemplo, sentem mais dificuldades no processo de aquisição/aprendizagem da língua (p. 1). Desta forma, conclui-se que quanto mais cedo estiver exposta uma criança a duas línguas, maior será o sucesso na sua aquisição de ambas.

Não obstante, a aquisição bilingue ou bilinguismo pode dar-se de diversas formas, em diferentes contextos e idades, desde e quando o falante tenha contacto regular com as duas (ou mais) línguas durante o período em que o conhecimento linguístico se esteja a construir (Almeida & Flores, 2017, p. 275). Muitos são os rótulos que caracterizam esta comunidade de falantes, que dependendo da sua proficiência e momento de aquisição das L1 e L2, podem ser bilingues recetivos, produtivos, precoces ou tardios (Bhatia, 2006, p. 18). Para o estudo em questão, centrou-se em dois tipos de aquisição bilingue: a aquisição simultânea (2L1) e a sucessiva (L2).

1.2.1. Aquisição simultânea (2L1) e sucessiva (L2)

A aquisição bilingue acontece quando uma criança é exposta a duas ou mais línguas, preferencialmente durante o seu PC, adquirindo esta um amplo conhecimento e domínio das estruturas linguísticas de ambas. Não obstante, a aquisição bilingue pode acontecer em diferentes

contextos socioculturais, por diversas causas e idades (Genesee, 2001, p. 165). Geralmente estas condicionantes, como o contexto ou a idade em que uma criança começa a ser exposta a duas ou mais línguas, têm um impacto no domínio e proficiência adquiridos pelo falante (Almeida & Flores, 2017, p. 275). No seu artigo sobre bilinguismo, Almeida e Flores apresentam possíveis cenários de como a aquisição de duas línguas pode acontecer.

Dentro da aquisição da linguagem existe a aquisição monolingue, onde um indivíduo é exposto a uma língua desde a nascença, sendo essa língua, a sua primeira língua ou língua materna (L1). Por outro lado, a aquisição bilingue pode ser (1) *simultânea* ou (2) *sucessiva*, como foi indicado na secção 1.2.. No primeiro tipo de aquisição, a criança é exposta de forma regular a duas línguas desde o nascimento, adquirindo, assim, duas línguas maternas (2L1) (Almeida & Flores, 2017; Meisel, 2011). Este tipo de aquisição é, também, conhecido como *Bilingual First Language Acquisition* (BFLA) (Almeida & Flores, 2017, p. 276). Almeida e Flores (2017) referem, ainda, que "o termo 'simultâneo' vem do facto de a primeira exposição às duas línguas ter ocorrido simultaneamente – por volta do nascimento, ou pouco tempo depois" (p. 276). Por outro lado, a aquisição sucessiva dá-se quando a introdução da segunda língua (L2) sucede de preferência até aos sete anos de idade (Bhatia, 2006, p. 18), sendo esta, também, conhecida como *Bilingual Second Language Acquisition* (BSLA) (Almeida & Flores, 2017, p. 276). É, também, comum um terceiro caso de aquisição de uma L2 tardia, ou em fase adulta. Este tipo de aquisição é frequente acontecer quando um indivíduo adquire uma L2 na escola, em contexto de sala de aula, ou, por exemplo, no caso em que tenha de emigrar por motivos profissionais (Bhatia, 2006, p. 16). Normalmente, quando este tipo de aquisição da L2 ocorre, dificilmente o falante adquire uma proficiência equiparada à de um nativo dessa língua (Almeida & Flores, 2017; Meisel, 2011). Nesse sentido, pode concluir-se que tanto uma aquisição 2L1 como a uma aquisição L2 infantil são mais semelhantes a uma aquisição L1 (monolingue) do que a uma aquisição L2 adulta/tardia.

Segundo Almeida & Flores (2017) a aquisição bilingue simultânea pode ocorrer quando (pp. 276-7):

- uma criança nasce num país (ou comunidade) onde se utiliza mais do que uma língua no quotidiano. É o caso, por exemplo, do Luxemburgo que tem três línguas oficiais (luxemburguês, alemão e francês), ou o caso do catalão na região da Catalunha, em Espanha. De notar que nestes casos, quando as línguas são partilhadas pela comunidade são ambas consideradas línguas maioritárias e possuem o mesmo nível de prestígio;

- a criança está inserida numa comunidade bilingue onde uma das línguas não é a língua oficial, sendo, por isso, denominada de língua minoritária. Estes casos são bastante comuns em contexto de migração, por exemplo, com a diáspora portuguesa residente em França ou na Alemanha, em que o português é considerado língua de imigração ou língua de herança, sendo, também, menos prestigiada que a língua do país de acolhimento;
- os pais da criança têm línguas maternas diferentes, sendo uma delas a língua do país de acolhimento e a outra minoritária. Normalmente, neste tipo de situações, o *input* linguístico numa das línguas pode ser ainda mais reduzido, caso não exista o suporte de uma comunidade no país que fale essa língua. Quando assim é, a aquisição dessa língua reduz-se às interações da criança com o seu cuidador no seio doméstico, designando-se, por isso, de *bilinguismo familiar*.

De acordo com as autoras, poder-se-á dar, ainda, o caso de uma criança crescer num ambiente bilingue ao estar exposta a uma L2 através de uma ama ou cuidador que não seja nenhum dos seus pais (Almeida & Flores, 2017, p. 278) ou, então, no caso dos pais terem duas línguas minoritárias diferentes da língua do país de acolhimento (Cruz-Ferreira, 2006 citado em Almeida & Flores, 2017).

1.2.2. Bilinguismo

Estima-se, assim, que metade da população mundial é bilingue e que dois terços das crianças no mundo crescem em ambientes bilingues (Crystal, 2003, p. 17). Não obstante, e seja o bilinguismo uma realidade bastante comum nos dias de hoje, é possível encontrar uma variedade de definições que, ancoradas em diversos fatores (como o contexto de aquisição, a idade, o grau de proficiência, entre outras), procuram definir este fenómeno. A comunidade linguística divide-se e as opiniões baseiam-se frequentemente nas experiências ou sentimentos dos falantes sobre a língua: o que elas transmitem *vs.* o que elas representam (Montrul, 2015, p. 4). Entenda-se, enquanto alguns investigadores salientam a integração cultural como o fator mais importante, por outro lado, muitos apontam que apenas um indivíduo com domínio equivalente de ambas línguas pode ser verdadeiramente considerado bilingue.

Em suma a esta falta de concórdia, o conceito de bilinguismo foi contestado durante várias décadas, uma vez que, nos séculos XIX e XX, a língua representava um símbolo identitário de um povo, associado à nação e que sustentava a ideologia desse então de "uma língua – uma nação"

(Lynch, 2017, p. 44). Ademais, o bilinguismo era percecionado como prejudicial para as crianças, porque poderia causar distúrbios a nível linguístico e cognitivo (Flores, 2019b, p. 278). Segundo esta crença, uma criança bilingue possuía menos capacidades linguísticas nas duas línguas, mesmo que uma delas fosse a sua língua materna, refletindo-se, por exemplo, no insucesso escolar. Apesar destas teorias pouco fundamentadas, Flores (2019b) afirma que até à data nenhum estudo linguístico conseguiu provar empiricamente que a exposição diária a duas línguas bloqueasse o processo de aquisição uma da outra (p. 278) e, consequentemente, o conhecimento e capacidades linguísticas do falante bilingue.

Mais tarde, com a globalização, a procura por melhores condições de vida e novas oportunidades, entre outros fatores, aumentaram, também, os fluxos migratórios decorridos ao longo de décadas. Estes, impulsionaram a mobilização de pessoas, que contribuiu para a diversificação e fusão de novas culturas e línguas numa sociedade cada vez mais multicultural. Como nos alerta Flores (2019b), apesar da larga existência de sociedades bilingues (e multilingues) presentes em países como a Índia ou Moçambique (p. 237), aumentaram, também, os números de crianças que foram crescendo em ambientes bilingues, embora fosse, na maioria das vezes, em contexto de migração, como é o caso, por exemplo, da comunidade hispano-falante residente nos EUA (Almeida & Flores, 2017, p. 277). Este paradigma contribuiu para que o bilinguismo fosse considerado um fenómeno mundial (Buchweitz & Prat, 2013, p. 429) e que, durante os anos 70, ocupasse a agenda dos linguistas, levando a que os mesmos começassem a estudar o fenómeno (Lynch, 2017, p. 46). Para Bhatia (2006) e Grosjean (1989) um falante bilingue demonstra possuir um conjunto de características e atributos que raramente são encontrados numa pessoa monolingue, e por essa razão, os autores rejeitam a ideia concebida por muitos de que um falante bilingue é a soma de dois monolingues numa só pessoa. Em oposição, Bloomfield (1933) argumenta que um falante bilingue é aquele que detém um controlo nativo de duas línguas, sugerindo, assim, que deve existir um equilíbrio perfeito ao nível da proficiência escrita e oral em ambas línguas – *'balanced bilingual'*. No entanto, encontrar falantes bilingues com este perfil é extremamente raro, pois os mesmos assumem perfis linguísticos bastante diversificados, que variam devido à relação do falante com a sua língua. Não obstante a estas definições, Almeida e Flores (2017) consideram que com o passar dos anos, o termo bilingue foi deixando de estar somente associado à ideia de proficiência linguística, passando este a designar falantes que possuem competências linguísticas em pelo menos duas línguas (p. 275). O contexto migratório é visto, por essa razão, o ambiente primordial e propício à aquisição de duas línguas, uma vez que possibilita que o falante esteja em

contacto com a língua de origem no seio familiar e com a língua do país de acolhimento, que é a língua falada pela sociedade dominante (Flores, 2017, p. 279).

1.2.3. Bilinguismo de herança

O termo língua de origem e/ou de herança (do inglês *heritage language*) é um termo bastante usado quando falamos sobre bilinguismo de herança. De acordo com Almeida e Flores (2017) "o termo 'língua de herança' (LH) designa uma língua adquirida desde a nascença, sobretudo em contexto familiar, mas que não é a língua dominante do falante bilingue" (p. 292). Relativamente ao seu contexto de aquisição, um dos cenários mais comuns é a aquisição bilingue em contexto migratório, onde o contacto com a LH incide, geralmente, no seio familiar. Não obstante, dependendo da representação da diáspora nesse país, o falante bilingue, pode, ainda, ter contacto com a sua LH caso exista uma expressiva comunidade do seu país de origem no país de acolhimento, como é o caso da diáspora portuguesa em França ou na Alemanha (Almeida & Flores, 2017). Nestes casos, o contacto com a LH é extensível, por exemplo, aos encontros e eventos organizados pela comunidade emigrante portuguesa em associações e clubes recreativos. Por seu turno, Montrul (2015) define LH como uma língua cultural ou etnolinguisticamente minoritária que se desenvolve num ambiente bilingue onde se fala uma outra língua socio politicamente maioritária (p. 2). A LH pode, ainda, ser designada como língua minoritária ou língua de imigração (Almeida & Flores, 2017, p. 277). Associado a esta, está, também, o termo 'falante de herança' (do inglês *heritage speaker*) que, segundo as autoras, teve origem no Canadá nos anos 70, para designar especificamente indivíduos de famílias imigrantes, que já nasceram no país de acolhimento ou que emigraram durante a infância, e que adquiriram a sua língua de origem no seio familiar, enquanto aprendiam a língua maioritária do país de emigração (p. 291). De notar que este grupo particular de falantes pode adquirir níveis de proficiência apenas parciais e bastante diversificados, uma vez que as experiências de aquisição e o contacto com a língua de origem podem variar (Flores & Melo-Pfeifer, 2014, p. 18). A LH está, assim, intrinsecamente aposta às raízes culturais e identitárias deste grupo de falantes, uma vez que a sua aquisição vai mais além dos aspetos linguísticos, existindo uma forte ligação e sentimento de pertença por parte do falante com o seu país e cultura de origem. Os estudos sobre falantes de herança tiveram início como parte da linguística de contacto e da sociolinguística, embora, mais recentemente, estes

falantes se tenham tornado um grupo importante na linguística experimental, particularmente nas áreas de aquisição e na psicolinguística (Benmamoun et al., 2013, p. 131).

Como mencionado acima, o falante de herança representa um grupo de falantes muito característico dentro do bilinguismo. Por conseguinte, a aquisição da LH é um tipo de aquisição bilingue precoce que tem lugar num ambiente sociolinguístico também muito específico (Montrul, 2015, p. 2). Este tipo de falantes bilingues são geralmente filhos de imigrantes que se estabeleceram num país diferente do seu país de origem, o qual é comummente designado neste tipo de estudos de país de acolhimento ou de emigração. O falante de herança, proveniente de famílias imigrantes que já nasceu no país de emigração ou que emigrou ainda na infância, é, então, emigrante de segunda ou terceira gerações (Almeida & Flores, 2017, p. 291). Este adquire, portanto, a sua LH em contexto familiar, uma vez que esta é a principal língua usada pela sua família – que representam a primeira geração de emigrantes. Conclui-se que o falante de herança é exposto à sua LH durante os primeiros anos de vida (até aos 3 anos de idade, senvivelmente), e só após a entrada no pré-escolar é que este começa a ter contacto com a língua do país de acolhimento, que se intensifica durante o seu crescimento (Almeida & Flores, 2017, p. 292). Por oposição, a LH é preferencialmente usada na comunicação diária com os pais, avós, tios ou outros imigrantes da mesma origem (Almeida & Flores, 2017, p. 292).

Nesta dissertação, as narrativas utilizadas como objeto de estudo foram escritas por crianças lusodescendentes residentes na Suíça (cantão alemão), que adquiriram o PE e o AP como línguas maternas, em contexto migratório. Estas crianças correspondem ao perfil acima descrito, uma vez que são filhos de imigrantes portugueses, em que muitos nasceram na Suíça ou emigraram durante a infância. Assim sendo, o contacto com o PE deu-se nos primeiros anos de vida, tendo o AP sido introduzido mais tarde, aquando da entrada no infantário, em muitos dos casos. O contacto com a língua do país de acolhimento, a língua maioritária, é, mais tarde, intensificado na escola e também através das relações e interação social estabelecidas fora o seio familiar pelo falante de herança (Almeida & Flores, 2017, p. 292). Almeida e Flores (2017) realçam, ainda, que o falante de herança, exposto às duas línguas desde idade precoce, desenvolve conhecimento nativo de dois sistemas linguísticos (p. 292), que são independentes.

Parte da investigação desenvolvida em torno da aquisição bilingue tem dado um grande destaque a este tipo de aquisição, em particular quando esta ocorre em contexto migratório (Flores & Melo-Pfeifer, 2014, p. 20). Flores & Melo-Pfeifer (2014) salientam que o *background* migratório tem

um papel fundamental por proporcionar o *input* a que a criança está exposta, permitindo que esta adquira as suas línguas a partir dessas evidências (p. 20). O facto é que se a aquisição da LH tiver apenas como suporte o meio familiar, esta pode desenvolver características muito próprias (Flores & Melo-Pfeifer, 2014, p. 21). Isto é, se determinadas estruturas linguísticas não existirem no *input* quotidiano a que a criança está exposta, por não serem tão usadas no registo oral, como é o caso, por exemplo, dos verbos no mais-que-perfeito do indicativo, a criança não irá adquirir este e outro tipo de estruturas linguísticas menos usadas na oralidade (Flores & Melo-Pfeifer, 2014, p. 20). O mesmo pode acontecer com a utilização de um vocabulário mais refinado ou mais específico, que caso a criança não esteja exposta a esse tipo de registo, o mais provável é que também não adquira esse tipo de conhecimento lexical. Quando assim sucede, é notório por parte dos falantes de herança que estão expostos apenas ao *input* doméstico, um maior à vontade com o registo coloquial e um menor domínio do registo formal. Por esse motivo, é aconselhável que os falantes de herança procurem infraestruturas fora do meio familiar, como aulas extracurriculares da língua de origem, que lhes permitam desenvolver um conhecimento linguístico mais rico e abrangente na sua LH (Flores, 2017, p. 281).

As particularidades do conhecimento linguístico de falantes de herança na sua língua de origem, muitas vezes considerado desviante, também se pode atribuir ao facto de, após a entrada no pré-escolar, a língua do país de acolhimento assumir uma presença mais forte por ser a língua mais utilizada no quotidiano, pela sociedade e, em muitos casos, por ser a língua em que o falante de herança se sente mais confiante em usar (Flores & Melo-Pfeifer, 2014, p. 19). Por oposição, a língua minoritária resume-se a um meio social mais restrito, uma vez que tampouco aufere de qualquer estatuto político no país de acolhimento. Este fator acentua o facto da amplitude lexical e o domínio de algumas estruturas gramaticais possa ser mais limitado nos falantes de herança (Flores & Melo-Pfeifer, 2014, pp. 39-40).

Para uma melhor análise do comportamento linguístico deste tipo de falantes, ter-se-á como objeto de estudo um corpus constituído por narrativas escritas por crianças lusodescendentes. No segundo capítulo, será feita uma introdução a este tipo de instrumento linguístico, às suas características e benefícios da sua utilização.

CAPÍTULO II – O CORPUS COMO RECURSO PARA ANÁLISE LINGUÍSTICA COM PARTICULAR ENFOQUE EM CORPUS DE APRENDIZAGEM

2.1. DEFINIÇÃO DE *CORPUS* E PRINCIPAIS CARACTERÍSTICAS

O termo *corpus* (cujo plural se designa de *corpora*) é amplamente utilizado em áreas de estudo da linguística, literatura, humanidades digitais, entre outras disciplinas que estudam matérias inerentes às línguas e ao seu processamento. Os linguistas, na sua definição mais geral, descrevem um corpus como um conjunto de textos escritos (ou excertos de textos) e transcrições de registos orais, geralmente utilizados para análise e estudos linguísticos (Fenlon & Hochgesang, 2022b; McEnery & Wilson, 2001; Mendes, 2016; Wallis, 2021). Porém, os avanços tecnológicos, em especial na área do PLN, trouxeram uma nova roupagem a este termo. De acordo com Hunston (2002a), o termo 'corpus' passou a ser utilizado para designar conjuntos de textos (ou partes de textos) que são armazenados e acedidos eletronicamente, dado que os computadores têm a capacidade de armazenar e processar grandes quantidades de informação (p. 2). Estes exemplares da língua natural, obedecem, por sua vez, a um conjunto de características que definem o seu objetivo. Fromm (2003) alerta que um corpus não deve ser equiparado a uma coletânea nem a uma antologia, sendo que este, de um modo geral, se refere a uma coleção de textos reunidos com um propósito específico de análise (p. 69). Por seu turno, e envergando uma perspetiva mais objetiva, Dash e Arulmozi (2018) apontam que um corpus constitui "a statistically sampled language database for the purpose of investigation, description, application and analysis relevant to all branches of linguistics" (p. 4). Os autores acrescentam que, devido à sua grande estrutura, composição variada, informação, autenticidade, ampla representação, facilidade de utilização e precisão, o uso dos corpora tornou-se um recurso indispensável em diferentes áreas da linguística (Dash & Arulmozi, 2018, p. 4).

A constante evolução tecnológica desencadeou, assim, a utilização e manuseamento deste tipo de recursos linguísticos no plano digital. A LC é uma das áreas da linguística que "baseia o estudo da língua em ocorrências extraídas de um corpus" (Mendes, 2016, p. 224). Apesar da longa e interessante história, este termo é relativamente moderno e descrito por McEnery e Wilson (2001) como "the study of language based on examples of 'real life' language use" (p. 1). Nesse sentido, atualmente, esta área da linguística depende fortemente de equipamentos eletrónicos que possibilitem a procura de formas ou estruturas linguísticas específicas através do processamento

eletrónico do corpus (Sampson & McCarthy, 2004, p. 1). Sampson e McCarthy (2004) referem que, consequentemente, a LC é, nos dias de hoje, percecionada como "linguística de corpus eletrónica" (p. 1), assim como o termo 'corpus' "is now almost synonymous with the term machine readable corpus" (McEnery & Wilson, 2001, p. 17). Segundo Mendes (2016) "para além do trabalho de compilação de dados, a Linguística de Corpus preocupa-se ainda com a anotação de informação linguística sobre os textos que compõem o corpus" (p. 225), tema que será alvo de uma discussão mais aprofundada no quarto capítulo. Entenda-se, portanto, que devido aos avanços tecnológicos, e consequentemente a uma adequação e nova conotação semântica, o termo corpus é, atualmente, utilizado para designar conjuntos de textos ou partes de texto armazenados e acedidos eletronicamente, que são, posteriormente, objeto de uma investigação específica e mais aprofundada.

Alguns investigadores (Fenlon & Hochgesang, 2022b; Gries & Berez, 2017; McEnery & Hardie, 2012; McEnery & Wilson, 2001; Mendes, 2016) consideram, assim, que um corpus deve possuir um certo conjunto de características, das quais, se destacam:

- Representatividade: Segundo Hunston (2022), "representativeness is about the relationship between the corpus and the population (of texts) it is taken from" (p. 33). Nesse sentido, o corpus destina-se a ser representativo, por exemplo, de um tipo específico de registo, orador, variedade ou da língua como um todo (Gries & Berez, 2017, p. 380). Por conseguinte, Sinclair (2004) afirma que "the larger and more representative the corpus the greater the attestation that is possible" (p. 4). Não obstante, é importante clarificar que um corpus representa essencialmente uma amostra e que um dos principais desafios consiste em garantir que essa amostra capte com exatidão a diversidade e as características dos informantes ou da língua que pretende representar. McEnery e Wilson (2001) alertam que analisar "every utterance in such a language would be an unending and impossible task", não só pela dimensão de textos existentes, mas também devido à continua evolução das línguas vivas (p. 30). No caso em que se pretenda analisar uma variedade inteira de uma língua em vez de um texto ou autor individual, só será possível, por exemplo, no caso de uma língua morta ou com poucos textos existentes (McEnery & Wilson, 2001, p. 29). Por conseguinte, Fenlon e Hochgesang (2022a) recomendam a constituição de um corpus com amostras selecionadas de acordo com critérios linguísticos específicos, que permitam que o perfil dos informantes sob análise seja o mais homogéneo possível (p. 3). Desta forma, quando um investigador descrever ocorrências de um

determinado fenómeno linguístico observado no seu corpus, poderá partir do princípio de que, com base na sua distribuição no corpus, haja uma distribuição semelhante numa comunidade mais alargada (Fenlon & Hochgesang, 2022a, p. 3). Por outro lado, Freitas (2015) realça a importância do reconhecimento do corpus como uma amostra, que é crucial para compreender as suas limitações e considerações em análises linguísticas (pp. 9-10).

- Equilíbrio: um conceito que está intrinsecamente associado à representatividade, é o de equilíbrio, que se refere às proporções das diferentes amostras incluídas num corpus (Ädel, 2020, p. 5). Segundo Hunstone (2022), "balance refers to the relative number of texts or tokens in each component of a corpus" (p. 31). Desta forma, o equilíbrio num corpus envolve a manutenção de uma distribuição proporcional e representativa de textos em diferentes componentes do mesmo. Também McEnery e Hardie (2011) acrescentam que num corpus equilibrado, "the relative sizes of each of [the subsections] have been chosen with the aim of adequately representing the range of language that exists in the population of texts being sampled" (p. 239). De acordo com os autores, através do equilíbrio, assegura-se que o corpus reflita uma diversidade de conteúdos linguísticos de forma a evitar enviesamentos ou a sobre representação de categorias específicas, conduzindo a uma análise mais fiável e abrangente.
- *Machine-readable*: o corpus é um recurso legível por máquina, isto é, deve estar disponível em formato digital, podendo ser pesquisado através de um computador. De acordo com McEnery e Wilson (2001), uma das vantagens de corpora legíveis por máquina é o facto de poderem ser pesquisados e manuseados de forma mais eficiente para obter resultados sobre a frequência das características linguísticas, o que não seria possível através de outro formato (p. 31). Ademais, uma vez em formato eletrónico, o corpus pode ser enriquecido com informação adicional (anotação) para efeitos de pesquisa linguística (McEnery & Wilson, 2001, p. 32). Essas anotações podem, pode sua vez, fornecer informações aos níveis fonológico, morfológico, sintático e discursivo (Fenlon & Hochgesang, 2022a, p. 3).
- Autêntico: por fim, a autenticidade é outra característica que se aplica aos corpora, uma vez que contém dados linguísticos recolhidos em ambientes naturais (Gries & Berez, 2017, p. 380). Isto significa que os dados linguísticos, ao serem recolhidos para integrar um corpus, não foram produzidos apenas com essa finalidade e, por conseguinte, essa característica permite que estes representem com uma maior precisão o uso quotidiano da língua (Fenlon & Hochgesang, 2022a, p. 4).

2.2. USABILIDADE DO CORPUS COMO FERRAMENTA DE PESQUISA

O estudo de corpora apresenta inúmeras vantagens. O uso de corpus para validar resultados de uma pesquisa é uma ferramenta utilizada há séculos, pese embora a ciência da LC seja relativamente recente (Fromm, 2003, p. 69). Como referido anteriormente, associado à utilização e estudo de corpora está a LC. Mendes (2016) explica que através da utilização de corpora é possível fundamentar análises linguísticas com base em contextos variados e num conjunto alargado de dados (p. 224).

Dash e Arulmozi (2018) apresentam alguns benefícios referentes à utilização de um corpus linguístico. Segundo os autores, um corpus representativo, bem planeado e equilibrado constitui um padrão empírico, que atua como uma referência para a validação do uso de propriedades linguísticas existentes numa língua (Dash & Arulmozi, 2018, p. 11). Com efeito, através de um corpus, pode obter-se *(a)* informação detalhada sobre todas as propriedades, elementos e componentes utilizados numa língua, *(b)* informação gramatical e funcional de palavras, frases, orações e expressões idiomáticas, *(c)* informação baseada na utilização de segmentos, morfemas, palavras, compostos, locuções e frases utilizadas numa língua, *(d)* indicações textuais e contextuais de um texto através do fornecimento de informações relativas ao tempo, lugar e agente de um evento linguístico e *(e)* informações extralinguísticas relacionadas com o discurso linguístico (Dash & Arulmozi, 2018, pp. 11-12). Segundo os autores, através da investigação aprofundada de um corpus podem, também, obter-se informações inerentes ao tecido social e cultural da comunidade linguística em análise (Dash & Arulmozi, 2018, p. 12). A exploração dos fatores que compõe o tecido sociocultural representado numa amostra de linguagem terão uma componente importante para a compreensão e análise das amostras recolhidas e exploradas neste projeto de dissertação.

2.3. TIPOLOGIA DE CORPUS

De acordo com Dash e Arulmozi (2018), "corpora can be of many types with regard to texts, languages, modes of data sampling, methods of corpus generation, manners of text processing, nature of text utilization, and so on" (p. 38). Desta forma, os corpora, para além de englobar uma grande variedade de textos, incluindo diferentes géneros, assuntos ou formatos, podem ser concebidos para diferentes línguas, divergir nos métodos utilizados para recolha e compilação das amostras, assim como variar no seu processamento e nos objetivos para os quais são utilizados.

Hunston (2002b) destaca, assim, a natureza multifacetada destes conjuntos de dados linguísticos, afirmando que os corpora revolucionaram o estudo da linguagem e das suas aplicações nas últimas décadas (p. 1). Os corpora linguísticos compreendem, assim, textos que, na sua forma original, podem ser escritos, falados ou gesticulados (no caso da linguagem gestual) (Hunston, 2022, p. 21). Dada a relevância para esta dissertação, far-se-á referência apenas aos corpora textuais e os corpora de fala.

1. *Text Corpus* ou **corpus textual** contém textos ou excertos de textos escritos, impressos, publicados ou provenientes de fonte eletrónica. Este tipo de corpus pode ser recolhido através de livros, jornais, revistas ou sítios da internet. Segundo Boulton (2017), o *Brown Corpus*[3] (Kučera & Francis, 1967) foi o primeiro corpus moderno de inglês americano, composto por um milhão de palavras recolhidas de 500 excertos de textos publicados em 1961 (p. 183). Nesse período, quando se deu início à criação de corpora eletrónico, a tecnologia informática disponível não era tão avançada como nos dias de hoje, o que dificultava a recolha e compilação de corpora através de um computador. Dash e Arulmozi (2018) acrescentam que "in those early years of electronic corpus generation, the Brown Corpus, which contained just 1 million words, was considered to be a standard database, since, at that particular time, a collection of 1 million words in electronic form was unthinkable for most linguists" (p. 20). De acordo com Boulton (2017), o objetivo da criação deste corpus "was to introduce greater scientific rigor from a more systematic base" (Boulton, 2017, p. 183). Mais tarde, os corpora de pequenas dimensões começaram a ser substituídos por corpora de grandes dimensões, devido aos avanços tecnológicos subsequentes dos computadores, que já permitiam o armazenamento e processamento de grandes volumes de texto (Dash & Arulmozi, 2018, p. 20). Assim, o corpus *Bank of English*[4] (BoE), um projeto internacional de corpus de língua inglesa, foi concebido para fins linguísticos e pedagógicos (Boulton, 2017, p. 183). O BoE representa uma enorme coleção eletrónica de amostras do inglês moderno, na sua versão escrita e falada, desenvolvido para análise da forma, significado, gramática e uso das palavras (Järvinen, 2003, p. 44). Foi lançado pela primeira vez em 1991 pelo *Collins Birmingham University International Language Database* (COBUILD) e pela Universidade de Birmingham, onde incluía cerca de 200 milhões de palavras (Järvinen, 2003, p. 43). O BoE é atualizado regularmente, fornecendo amostras objetivas do uso da língua inglesa no quotidiano. Muitas

[3] http://icame.uib.no/brown/bcm.html
[4] https://cqpweb.bham.ac.uk/

destas primeiras referências internacionais de corpora textuais foram produzidas para criar dicionários e enriquecer os já existentes. Abeillé (2003) explica, por exemplo, que "the main motivation for the Bank of English project was to extend the Collins COBUILD dictionary, with new examples and new constructions for most of the English verbs". Também o *British National Corpus*[5] (BNC) é composto por uma coleção de mais de 100 milhões de palavras de amostras representativas de linguagem, tanto numa versão escrita como falada (Burnard, 2009). Segundo Abeillé (2003), o BNC foi dos primeiros corpora a ser construído para o inglês, tendo, por isso, contribuído para o desenvolvimento de ferramentas de PLN em inglês e da própria linguística de corpus inglesa. Este corpus foi cientificamente concebido para representar uma ampla secção transversal do inglês britânico, tendo a sua construção começado em 1991, ficando concluído em 1994 (Burnard, 2009). Não foram acrescentados novos textos após a conclusão do projeto, mas o corpus foi ligeiramente revisto antes do lançamento da segunda edição – *BNC World* (2001) – e da terceira edição – *BNC XML Edition* (2007) (Burnard, 2009). O projeto foi liderado pela *Oxford University Press*, com a participação da maioria dos editores de dicionários de Inglaterra (Burnard, 2009). A parte escrita do BNC (90%) inclui, por exemplo, extratos de jornais regionais e nacionais, periódicos e revistas especializadas para todas as idades e interesses, livros académicos e de ficção popular, cartas e memorandos publicados e não publicados, ensaios escolares e universitários, entre muitos outros tipos de texto (Burnard, 2009). De acordo com Taylor e Barker (2008), no final da década de 1990, para além das coleções de inglês britânico, estavam a ser criados corpora importantes de outras variedades, como o *American National Corpus*[6] (ANC) (p. 243), que foi concebido e desenvolvido com o objetivo de reunir uma base de dados linguística representativa do inglês americano contemporâneo, que pudesse ser usada para contribuir para a investigação linguística, bem como para fornecer um recurso para o ensino (Reppen & Ide, 2004, pp. 105-107). A análise do BNC demonstrou que, devido a várias diferenças na linguagem utilizada nos dois países, o BNC não podia ser utilizado como um recurso de referência para estudar a variedade do inglês americano. Desta forma, o objetivo do ANC centrou-se em obter uma grande base de dados de, pelo menos, 100 milhões de palavras, que se tornou suficientemente comparável em todos os géneros com o BNC. O corpus foi desenvolvido com a contribuição de um consórcio de editores de dicionários de inglês americano e com a contribuição de

[5] http://www.natcorp.ox.ac.uk/
[6] https://anc.org/

empresas com interesses na área da língua e da linguística (Reppen & Ide, 2004, p. 105). Uma vez que o projeto está em processo de desenvolvimento, apenas a primeira parte, que compreende uma base de dados de 10 milhões de palavras, foi, até agora, disponibilizada para acesso público.

No PE, o *Corpus de Referência do Português Contemporâneo*[7] (CRPC) representa um dos mais vastos e mais diversificados corpus de português europeu disponível *online*, que tem vindo a ser desenvolvido pelo Centro de Linguística da Universidade de Lisboa (CLUL) desde 1988 (Nascimento et al., 2014, p. 237). O CRPC constitui um recurso de "exploração de corpora essencial à comunidade científica para estudos nas áreas da linguística e do PLN" (Mendes et al., 2012, p. 466). A constituição deste corpus eletrónico teve como objetivo "fornecer informação abrangente sobre o português contemporâneo escrito e oral" (Mendes et al., 2012, p. 467), refletindo tanto as variedades regionais e nacionais do PE, onde estão incluídas amostras do português de Angola, Cabo Verde, Guiné-Bissau, Moçambique, São Tomé e Príncipe, Goa, Macau, Timor-Leste e do Brasil (Nascimento et al., 2014, p. 238). Nascimento *et al.* (2014) referem, ainda, que "the CRPC currently contains approximately 310 million words, searchable through a user-friendly interface, and it is envisaged as a monitor corpus (from which one can extract balanced subcorpora) that can serve as a sample of the Portuguese language" (p. 237). A plataforma oferece uma variedade de opções de pesquisa a vários níveis, permitindo aos utilizadores o manuseamento do corpus de acordo com os seus objetivos de investigação (Généreux et al., 2012, p. 114).

Por seu turno, um exemplo de um corpus textual no AP é o corpus DEREKO (em alemão *Deutsches Referenzkorpus*), conhecido em inglês, também, como o *Archive of General Reference Corpora of Contemporary Written German* (ISD, 2023). Segundo Kupietz *et al.* (2010), o corpus DEREKO "is one of the major resources worldwide for the study of the German language" (p. 1848), e em janeiro de 2020 contava com mais de 46,9 mil milhões de palavras (ISD, 2023). O corpus foi criado em 1964 com o principal objetivo de proporcionar uma base empírica para a investigação linguística (Kupietz et al., 2010, p. 1848). Contém textos de ficção, científicos e de jornais, assim como de outras fontes, e pode ser acedido gratuitamente através das plataformas COSMAS II[8] e KorAP[9] (Kupietz et al., 2010, p. 1848).

[7] http://gamma.clul.ul.pt/CQPweb/
[8] https://www2.ids-mannheim.de/cosmas2/
[9] https://korap.ids-mannheim.de/

2. Por seu turno, os **corpora de fala**, segundo Gut (2020), "contain spoken language, usually referred to as *speech corpora* (or speech databases) and *spoken corpora* respectively" (p. 236). Na prática, o que diferencia estas duas variantes de corpora de fala são, primeiramente, a grande quantidade de amostras de linguagem falada existente nos '*speech corpora*', geralmente recolhida em condições experimentais e, em segundo lugar, as amostras são utilizadas para aplicações industriais e tecnológicas, como por exemplo na avaliação de sistemas de reconhecimento automático da fala (Gut, 2020, p. 237). Em contrapartida, os '*spoken corpora*' são compilados para fins linguísticos, como no estudo do uso da língua e da comunicação humana e em aplicações linguísticas, como no ensino das línguas e na elaboração de gramáticas e dicionários (Gut, 2020, p. 237). Por outro lado, Dash e Arulmozi (2018) classificam os '*spoken corpora*' como uma extensão dos '*speech corpora*' (p. 35). Enquanto os '*speech corpora*', para os autores, contêm amostras de texto obtidas a partir de interações verbais (textos que estão disponíveis em formato áudio), os '*spoken corpora*' diferenciam-se por armazenar "the transcripted version of audio texts collected as speech corpus" (p. 35). Dash e Arulmozi (2018) acrescentam que "a Spoken Corpus is a different class of its own which has properties of both the text and the speech corpus but has evolved into a new type of corpus due to its nature of the composition" (p. 35). Não obstante, os corpora de fala tanto podem conter amostras da linguagem cuja produção se obteve de origem espontânea, ou seja, sem a intervenção do investigador, ou estes podem, ainda, conter amostras que foram intencionalmente produzidas e controladas pelo investigador. Geralmente, as amostras presentes nos corpora de fala versam sobre interações do quotidiano, como conversas, telefonemas, discursos, entre outros tipos (Dash & Arulmozi, 2018, p. 42). Por se tratarem de amostras da linguagem produzidas através do registo oral, possibilita a exploração e análise de outros fenómenos linguísticos, extraídos, por exemplo, através das anotações fonética, fonológica e prosódica (Gut, 2020, pp. 241-242). O corpus *The London-Lund Corpus of Spoken English*[10] (LLC), iniciado por Jan Svartvik na Universidade de Lund (Suécia) em 1975, representa uma das mais importantes referências nos corpora de fala por se tratar do primeiro corpus de fala aberto legível por máquina a ser compilado (Põldvere et al., 2021, p. 459). Nesse sentido, o LLC representa um marco importante na elaboração de um corpus de fala, na medida em que estabeleceu alguns critérios que viriam a servir de referência para a compilação de outros corpora de fala. O LLC é constituído por transcrições de registos orais

[10] https://varieng.helsinki.fi/CoRD/corpora/LLC/

(*spoken texts*) provenientes de dois projetos: a primeira parte foi feita com dados retirados do *Survey of English Usage* e a segunda parte foi feita com dados retirados do *Survey of Spoken English* (Svartvik, 1990, p. 11). Segundo Francis (1992), "the entire corpus contains 200 texts of 5000 words each, half of which were recorded from various types of spoken English, some of it public (e.g. radio programs) and some of it private" (p. 197) e foi gravado com falantes adultos instruídos desde os anos de 1950 até 1980 (Põldvere et al., 2021, p. 461). A compilação do LLC teve como principal objetivo fornecer uma base suficientemente abrangente para o estudo da variação gramatical e estilística do inglês britânico falado (Svartvik, 1990, p. 11).

No PE, um exemplo de um corpus de fala é o *Perfil Sociolinguístico da Fala Bracarense*[11] (PSFB), do Centro de Estudos Humanísticos da Universidade do Minho, que constitui a primeira base sociolinguisticamente controlada de dados de fala em Portugal, num total de 90 falantes entrevistados, onde cada entrevista tem a duração de uma hora e versa sobre assuntos diferenciados (Herdeiro & Barbosa, 2015, pp. 334). Existe, ainda, o corpus *Santos – Português Europeu*[12] (Santos, 2006), que é um corpus de fala de crianças e de fala dirigida a crianças transcrito de acordo com o formato definido no projeto *Child Language Data Exchange System*[13] (CHILDES) (Santos et al., 2014, p. 1488). Segundo Santos *et al.* (2014), "the data were collected using videotape and correspond to child-adult interaction in a naturalistic setting: children were taped at their homes interacting with their family (most often their mother) and the researcher" (p. 1488). Uma vez que os dados se destinavam a servir a investigação sobre a sintaxe e a interface sintaxe-discurso, todos os enunciados de adultos e crianças foram transcritos (Santos et al., 2014, p. 1488). O corpus inicial foi alargado com mais amostras, assim como com novos recursos, nomeadamente o alinhamento e a etiquetação da transcrição sonora (p. 1489).

Em AP, o corpus *Forschungs- und Lehrkorpus Gesprochenes Deutsch*[14] (FOLK) consiste num corpus aberto com cerca de 300 horas de gravações áudio e vídeo de alemão falado em vários tipos de interação, com cerca de 3 milhões de *tokens* transcritos (Schmidt, 2016, p. 397). Desta forma, o corpus abrange uma vasta gama de tipos de interação em contextos variados (privados, institucionais e públicos), é diversificado e encontra-se transcrito, anotado e acessível

[11] https://sites.google.com/site/projectofalabracarense/
[12] O corpus está disponível na base de dados CHILDES, através da ligação: https://childes.talkbank.org/data/Romance/Portuguese/
[13] https://childes.talkbank.org/
[14] https://agd.ids-mannheim.de/folk.shtml

para a comunidade científica numa base de dados sem restrições de utilização (Schmidt, 2016, p. 397). O corpus *German political speeches from 21st century*[15] é constituído por discursos políticos alemães proferidos por altos funcionários da República Federal Alemã, sobretudo a partir de 1990, selecionados de acordo com a sua relevância política (Barbaresi, 2018, p. 793). O corpus reúne quatro tipos de oradores, que correspondem atualmente às quatro funções mais elevadas do Estado federal alemão, onde estão, ainda, incluídos discursos de ministros e secretários de Estado (Barbaresi, 2018, 793). Os dados foram recolhidos entre 2012 e 2017 e contabilizam um total de aproximadamente 10,9 milhões de *tokens* (Barbaresi, 2018, p. 792). O corpus encontra-se disponível em formato XML, tendo sido tokenizado e etiquetado ao nível morfossintático (Barbaresi, 2018, p. 796) e pode ser consultado através de uma pesquisa de texto integral[16] com anotação linguística. Outro exemplo no AP é o *Das KiezDeutsch-Korpus*[17] (KidKo), que contém a gravação de conversas maioritariamente em alemão, entre adolescentes com idades compreendidas entre os 14 e os 17 anos de idade (Wiese et al., 2012, pp. 102-3). Este projeto aborda a utilização de uma nova linguagem juvenil proveniente dos bairros urbanos multiétnicos de Berlim, o chamado "Kiezdeutsch", e onde são exploradas análises linguísticas com base neste recurso (Wiese, 2011, p. 146). Wiese (2011) denota que, contrariamente à perceção generalizada, o "Kiezdeutsch" não é um "gebrochenes Deutsch", mas sim um novo dialeto urbano do alemão que se desenvolveu na vida quotidiana de jovens, independentemente de terem ou não antecedentes migratórios (p. 147). O corpus é constituído por cerca de 333.000 *tokens* e a recolha de dados teve lugar em 2008 e compreende, principalmente, comunicações presenciais, mas também algumas conversas telefónicas, constituindo, portanto, momentos de conversação orais, informais e de interação (Wiese et al., 2012, p. 103). As 66 horas de gravações que compõe o corpus foram transcritas e encontram-se disponíveis em formato XML, estando alinhadas com os ficheiros de áudio (Wiese et al., 2012, p. 104). Os dados estão, assim, disponíveis em formato de texto, mas que permite a notação da língua falada, características específicas da variedade e individuais, bem como informação prosódica (Wiese et al., 2012, p. 104).

Como constatado, a comparação entre corpora de texto e corpora de fala realça as diversas formas como a língua pode ser captada e estudada. Enquanto os corpora de texto oferecem uma fonte rica de linguagem escrita, por outro lado, os corpora de fala fornecem informações sobre as

[15] https://politische-reden.eu/
[16] https://www.dwds.de/r?corpus=politische_reden
[17] https://www.linguistik.hu-berlin.de/de/institut/professuren/multilinguale-kontexte/korpora/kiezdeutschkorpus

nuances presentes na língua falada, seja esta em contextos mais ou menos formais. No entanto, a exploração deste tipo de recurso linguístico está longe de se resumir apenas à sua forma escrita ou falada, existindo, portanto, uma vasta classificação dos corpora segundo vários critérios como a sua extensão, domínio (geral ou especializado) ou quanto ao número de línguas que albergam (monolingue ou bilingue), por exemplo. Através do seu livro, Dash e Arulmozi (2018) oferecem uma clara e concisa abordagem sobre a diversidade de corpora, onde explicam com detalhe os vários subconjuntos que existem. Não obstante, um dos subconjuntos relevantes a esta dissertação, no domínio da análise de corpora, são os corpora de aprendizagem. Estes, que também podem albergar dados linguísticos na sua vertente escrita ou falada, destacam-se por constituírem um recurso fundamental para o estudo da aquisição de uma segunda língua (Granger et al., 2015, p. 1). Os corpora de aprendizagem compilam, desta forma, dados linguísticos que permitem aos investigadores aprofundar fenómenos linguísticos particulares, realçando a evolução das competências linguísticas e os padrões que surgem durante o processo de aprendizagem de línguas.

2.4. CORPUS DE APRENDIZAGEM

Como o próprio nome sugere, o corpus de aprendizagem, em inglês '*learner corpus*', compila um conjunto de dados linguísticos produzidos por aprendentes de uma língua (Hunston, 2002, p. 15). Numa vertente mais empírica, Gilquin (2020) oferece uma definição dos corpora de aprendizagem que vai de encontro ao objeto de estudo desta dissertação, onde afirma que:

> Typically, the term [learner corpora] covers both foreign language and second language situations, that is, respectively, situations in which the target language has no official function in the country and is essentially confined to the classroom (and, possibly, international communication), and situations in which the target language is learned by immigrants in a country where it is the dominant native language. (p. 283)

Os corpora de aprendizagem englobam, assim, dados linguísticos (escritos ou orais) de falantes que estão a adquirir uma segunda língua, o que pode ocorrer tanto em situação de emigração (onde a língua em aquisição não é a língua oficial do país), como em situações de segunda língua, no caso de se tratar da língua dominante do país de acolhimento. Segundo Granger *et al.* (2015), o aparecimento dos corpora de aprendizagem surgiu da necessidade de produzir ferramentas pedagógicas mais conscientes e centradas no informante (p. 1). Para além disso, referem os

autores, durante muito tempo, os dados utilizados para a criação de corpora de aprendizagem eram resultantes de tarefas linguísticas altamente controladas pelos investigadores e, por conseguinte, não refletiam necessariamente a maneira como os informantes comunicavam em contextos mais naturais (Granger et al., 2015, p. 1). Desta forma, Granger *et al.* (2015) classificam este tipo de corpus "as electronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria" (p. 1). Também Lozano *et al.* (2021) afirmam que "Learner corpus data is not necessarily authentic data, i.e., uncontrolled language produced for communicative purposes in natural settings, as it often consists of language samples produced in classroom contexts and elicited through particular linguistic tasks" (p. 138). Embora não reflitam na sua grande maioria situações de comunicação espontânea, os dados linguísticos obtidos para a compilação de um corpus de aprendizagem são valiosos porque são contextualmente ricos, mesmo que consistam, frequentemente, em amostras de linguagem geradas em ambientes estruturados, como em contexto de sala de aula, e obtidas através de tarefas linguísticas específicas concebidas para suscitar determinados comportamentos ou respostas, como é o caso do objeto de estudo desta dissertação.

2.4.1. Investigação com corpus de aprendizagem

De acordo com Gries e Berez (2017), nos últimos 15 anos assistiu-se a um rápido crescimento nas investigações linguísticas baseadas em corpora sobre o uso de línguas não nativas por falantes L2 e L3 (p. 391). Só no início da década de 1990 é que editores e académicos – de forma simultânea, mas independente – começaram a recolher e a analisar dados de aprendentes (Granger, 2003, p. 538). Esse crescimento foi facilitado por uma variedade de projetos de compilação de corpus deste tipo, com destaque para o *International Corpus of Learner English*[18] (ICLE) (Gries & Berez, 2017, p. 391). O ICLE foi o primeiro corpus de aprendizagem de grande escala a ser compilado, desenvolvido na Universidade de Lovaina, na Bélgica, por Sylviane Granger e outros investigadores (Lozano et al., 2021, p. 139). Inicialmente, o corpus continha cerca de 2 milhões de palavras e era constituído por pequenos textos argumentativos escritos por aprendentes de inglês (nível avançado) de diferentes origens linguísticas e foi posteriormente alargado e complementado por vários outros corpora (Lozano et al., 2021, p. 139). Granger (2003) afirma que o ICLE constitui um corpus "muito bem documentado", dado que para além

[18] https://corpora.uclouvain.be/cecl/icle/trial/

dos dados recolhidos, foram registadas e tidas em consideração mais de 20 variáveis relativas ao perfil sociolinguístico dos participantes, obtidas através de questionários preenchidos pelos mesmos (p. 539). Algumas dessas variáveis, que se encontram armazenadas numa base de dados, compreendem, por exemplo, o meio, a idade, o género, ou a língua de origem dos informantes (Granger, 2003, p. 539). À semelhança do ICLE, também o corpus em análise nesta dissertação apresenta características muito próximas, pois os dados sociolinguísticos recolhidos durante o período de testagem serão tidos em consideração e abordados no quinto capítulo.

De acordo com Mendes *et al.* (2016), a importância de dados empíricos obtidos através de corpora de aprendizagem tem sido cada vez mais reconhecida para estudos de Aquisição de Segunda Língua (em inglês *Second Language Acquisition*) e ensino/aprendizagem de línguas, embora estejam longe de atingir o seu impacto potencial, em parte, sublinham as autoras, devido à falta de recursos para as línguas além do inglês (Mendes et al., 2016, p. 3207). A limitação dos estudos com corpora de aprendizagem na língua inglesa era compreensível, refere Granger *et al.* (2015), tendo em conta a posição do inglês como a principal língua franca a nível internacional (p. 2). Não obstante, a transformação das sociedades em grupos mais diversos e multilingues, contribuiu para a mudança de paradigma (Granger et al., 2015, p. 2). Gilquin (2020) refere que gradualmente, foram aparecendo corpora de aprendizagem noutras línguas, tanto na sua versão escrita como falada, notando-se, porém, uma crescente recorrência à compilação de corpora de aprendizagem escrito diretamente a partir de fontes eletrónicas (p. 284). O autor acrescenta, ainda, que esse fenómeno facilita a compilação dos corpora (Gilquin, 2020, p. 284). O sítio da internet *Learner Corpora Around the World*[19], mantido pela Universidade de Lovaina, contém atualmente cerca de 200 corpora de aprendizagem eletrónicos, que compilam dados contínuos escritos e/ou falados produzidos por aprendentes de línguas estrangeiras ou falantes L2. Grande parte dos mesmo incidem sobre a língua inglesa, pese embora haja um número significativo de corpora de aprendizagem que estude outras línguas, como o árabe, francês, alemão, espanhol, português, entre outras.

2.4.2. Projetos de criação de corpus de aprendizagem de referência

No seu artigo *Linguística de Corpus e outros usos dos corpora em linguística*, Mendes (2016) apresenta algumas iniciativas de corpora de aquisição e de aprendizagem em PE. A disposição de

[19] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

dados extraídos de corpora para análise do português como língua primeira e língua segunda são fundamentais, segundo a autora (Mendes, 2016, p. 234). Relativamente à sua aquisição, Mendes (2016) apresenta o corpus *Base de Dados de Aquisição do Português*[20] (AcEP) com dados longitudinais espontâneos de crianças portuguesas monolingues e crianças bilingues luso-francesas, com idades entre os 0 e os 4 anos (p. 234). O projeto teve como objetivos o enriquecimento e disponibilização de recursos relacionados aos dados de produção oral por crianças portuguesas na aquisição do PE; a descrição linguística dos constituintes fonológicos, nomeadamente dos aspetos *segmento*, *sílaba* e *acento de palavra*; a contribuição para a discussão sobre a arquitetura do conhecimento fonológico e morfossintático no sistema linguístico das crianças e a produção de materiais para investigadores na área da aquisição da linguagem (CLUL, 2019a). Os dados foram recolhidos entre 1990 e 2000 e encontram-se em formato vídeo e áudio (CLUL, 2019a). Este projeto teve início em 1992 no Laboratório de Psicolinguística da Faculdade de Letras da Universidade de Lisboa, no âmbito do projeto de doutoramento apresentado por M. João Freitas (Freitas, 1997). A recolha de dados foi feita por vários investigadores, que ficaram responsáveis por um subconjunto de dados registados, perfazendo um total de 9 subconjuntos (CLUL, 2019a).

À semelhança do supramencionado *Corpus Santos* (Santos, 2006), também o *Corpus Batoréo*[21] (Batoréo, 2000) está disponível no projeto CHILDES. Este corpus de narrativas em português foi recolhido no âmbito do projeto da tese de doutoramento em Linguística por Hanna Jakubowicz Batoréo e centra-se no conhecimento linguístico do espaço (CHILDES, 2000). *Corpus Batoréo* constitui um corpus transversal de produções de narrativas induzidas a partir de duas sequências de imagens contadas por 30 adultos e 30 crianças, entre 1992 e 1993 (Mendes, 2016, p. 235). As narrativas foram produzidas por 10 crianças de 5 anos, 10 crianças de 7 anos, 10 crianças de 10 anos e um grupo de controlo de 30 adultos (Batoréo, 2000, p. 580).

O *The Learner Corpus of Portuguese as Second/Foreign Language*[22] (COPLE2) é um corpus de aprendizagem que inclui materiais escritos e falados e que visa fornecer dados empíricos para o ensino e aprendizagem do português como L2 ou língua estrangeira (LE), proporcionando, ainda, um recurso para o desenvolvimento de estudos interlinguísticos e materiais didáticos para a investigação da aquisição de uma segunda língua (Mendes et al., 2016, p. 3207). A subparte

[20] https://www.clul.ulisboa.pt/recurso/acquisition-european-portuguese-databank
[21] https://childes.talkbank.org/access/Romance/Portuguese/Batoreo.html
[22] http://teitok.clul.ul.pt/cople2/

escrita contém 966 ensaios, constituindo um total de 156.691 *tokens*, produzidos por 424 estudantes entre 2010 e 2012 (Mendes et al., 2016, p. 3207). Os 483 alunos têm entre 18 e 40 anos de idade e representam 14 línguas maternas diferentes (Mendes et al., 2016, p. 3208). O corpus foi, por sua vez, digitalizado, transcrito e guardado em formato XML, para ser tokenizado, etiquetado morfossintaticamente e lematizado (Mendes et al., 2016, p. 3209).

Mendes *et al.* (2016) fazem, também, referência a três corpora de aprendizagem do português. O corpus *Recolha de dados de Aprendizagem do Português Língua Estrangeira*[23], que contém 470 textos realizados por 397 informantes, falantes de 28 diferentes línguas maternas, é composto por 70.500 *tokens* (CLUL, 2019b). O corpus teve como principal objetivo recolher produções de aprendentes de português como LE com vista à criação de uma base de dados que possa apoiar a investigação na área da língua portuguesa (CLUL, 2019b). O corpus de *Produções Escritas de Aprendentes de PL2*[24] (PEAPL2), coordenado por Cristina Martins (C. Martins, 2013), integra, atualmente, três subcorpora pesquisáveis de movo autónomo, nomeadamente o subcorpus *Português Língua Estrangeira*[25] (PEAPL2_PLE), subcorpus *Timor*[26] (PEAPL2_Timor) e subcorpus *Guiné-Bissau*[27] (PEAPL2_Guiné-Bissau) (Ferreira et al., 2023). O projeto teve como principal objetivo a recolha de textos produzidos por aprendentes não nativos, a fim de possibilitar o apoio à investigação em aquisição/aprendizagem da língua estrangeira, neste caso o português, bem como a formação de professores e a produção de materiais didáticos (Araújo & Trabulo, 2014, pp. 11-12). Por fim, o *Corpus de Aquisição de L2*[28] (CAL2), que contém produções de adultos e crianças, num total de 281.301 palavras (Mendes et al. 2016, p. 3207), reúne os dados de produção espontânea (escritos e orais) recolhidos no âmbito do projeto Morfologia e Sintaxe na Aquisição de L2[29] (Ferreira et al., 2023, p. 249). O projeto CAL2 está disponível mediante registo prévio.

O *L1 Português – L2 Espanhol* é um subcorpus do *Corpus Escrito del Español L2*[30] (CEDEL2), que representa um recurso metodológico para a investigação de uma vasta gama de tópicos sobre aquisição de segunda língua, acessível *online* e gratuitamente (Lozano et al., 2021, p. 137). O

[23] https://www.clul.ulisboa.pt/recurso/recolha-de-dados-de-ple
[24] https://teitok2.iltec.pt/peapl2/#http://teitok.iltec.pt/peapl2/
[25] https://teitok2.iltec.pt/peapl2-ple/index.php?action=home
[26] https://teitok2.iltec.pt/peapl2-timor/index.php?action=home
[27] https://teitok2.iltec.pt/peapl2-gb/index.php?action=home
[28] https://cal2.clunl.fcsh.unl.pt/
[29] https://clunl.fcsh.unl.pt/projetos/projetos-concluidos/morfologia-sintaxe-na-aquisicao-l2/
[30] http://cedel2.learnercorpora.com/

corpus contém 21.662 palavras escritas por 164 participantes, com idades entre os 13 e os 53 anos, sendo que 98% dos participantes têm o PE como única L1 (Lozano et al., 2021, p. 146).

O *Corpus of Learner German* (CLEG) é um corpus que compila redações de escrita argumentativa em forma de resumos, ensaios e críticas redigidas por estudantes universitários britânicos, tendo o alemão como sua L2 (Maden-Weinberger, 2015, p. 34). O corpus foi compilado entre 2003 e 2005, ao longo dos anos académicos de três grupos de alunos com idades compreendidas entre os 18 e os 24 anos (Maden-Weinberger, 2015, p. 35). O estudo procurou analisar o uso do condicional (em alemão '*Konjunktiv*') em falantes L2 de alemão de diferentes níveis de proficiência, comparando-os com dados de falantes nativos (Maden-Weinberger, 2015, p. 25). Os informantes tiveram entre cinco e sete anos de aulas de alemão da escola secundária e os textos foram escritos em contexto de sala de aula, sem acesso a materiais de auxílio (Maden-Weinberger, 2015, p. 35).

À semelhança do CLEG, o corpus *Czech as a Second Language with Spelling, Grammar and Tags*[31] (CzeSL-SGT) também inclui transcrições de ensaios escritos por falantes não nativos da língua checa, recolhidos em 2013 (Rosen, 2014, p. 1). Segundo Rosen (2014), "word forms are tagged by word class, morphological categories and base forms (lemmas) (p. 1). Os autores do corpus procederam à verificação dos originais e correção de alguns erros, que foram, posteriormente, etiquetados com etiquetas de erro (Rosen, 2014, p. 1). O corpus está disponível para pesquisa *online* utilizando a interface de pesquisa do *Corpus Nacional Checo*[32] (Rosen, 2014, p. 1).

Como já foi mencionado ao longo deste capítulo, o projeto CHILDES consiste num vasto repositório, acessível através da internet, que disponibiliza corpora de linguagem infantil, recolhido por vários investigadores (Sokolov & Snow, 1997, p. 654). Este, assim como outros projetos, como o *TalkBank*[33] (MacWhinney, 2019), utilizam um sistema normalizado para a produção de transcrições informatizadas, denominado de CHAT (em inglês *Codes for the Human Analysis of Transcripts*) (MacWhinney, 1992, p. 14). Segundo MacWhinney (1992), "the system provides options for basic discourse transcription as well as detailed phonological and morphological analysis" (p. 14). Para além disso, os dados linguísticos transcritos de acordo com o sistema CHAT, podem, também, ser analisados pelos programas CLAN[34] (em inglês *Computerized*

[31] http://utkl.ff.cuni.cz/learncorp/
[32] https://www.korpus.cz/
[33] https://www.talkbank.org/
[34] "CLAN is designed specifically to analyze data transcribed in the CHAT format" (MacWhinney, 2019, p. 1921).

Language Analysis) (MacWhinney, 1992, p. 14). A criação desta base de dados eletrónica teve início em 1984 por Brian MacWhinney e Catherine Snow, com o intuito de auxiliar investigações no âmbito da linguagem infantil (Weissenborn, 1989, p. 259). Atualmente a plataforma CHILDES contém milhares de dados, disponíveis em vários formatos, dos quais tratam temáticas como distúrbios da linguagem, aquisição de linguagem nos níveis sintático, semântico, fonológico, morfológico e lexical, bem como temas inerentes ao bilinguismo, estudos interlinguísticos e a aquisição de uma segunda língua (Higginson, 1990, p. 474).

2.4.3. Desafios de processamento do corpus de aprendizagem

No que diz respeito à exploração de um corpus de aprendizagem, existem vários desafios associados dos quais se destaca, por exemplo, a anotação, em particular às tentativas de anotação automática (Gries & Berez, 2017, p. 391). Embora a compilação de corpora siga critérios expecíficos, Granger (2003) sublinha "extra care has to be taken in collecting the data for learner corpora given the large number of variables affecting the learning/acquisition process" (p. 538). Gries e Berez (2017) apontam que a probabilidade de encontrar, por exemplo, erros ortográficos, itens e estruturas gramaticais consideradas 'não-nativas' é mais elevada em corpora de aprendizagem do que noutros tipos de corpora (p. 391). Esta característica, por sua vez, dificulta a anotação automática do corpus por ferramentas, como etiquetadores ou lematizadores treinados. Com efeito, Gries e Berez (2017) apontam "Thus, such annotation efforts will likely require great care in choosing the right tagset and tagging algorithm, and more manual checking than is customary for native language use" (p. 391), a fim de contornar possíveis obstáculos durante o processamento do corpus. Não obstante, e devido à abundancia desse tipo de elementos em corpora de aprendizagem, "error tagging has been the most frequent type of annotation used in learner corpus research up to now" (van Rooy, 2015, p. 79). Van Rooy (2015) acrescenta "a substantial amount of research is directed at the development of automatic tools for the annotation of data, targeting linguistic properties such as word class, syntactic structure or semantic fields" (p. 79). Desta forma, a exploração mais recorrente nos corpora de aprendizagem incide na anotação de erros, sejam ortográficos ou, por exemplo, na identificação da utilização de expressões linguísticas 'não-nativas' presentes no corpus (Gries & Berez, 2017, p. 391).

Contudo, o processamento levado a cabo para a exploração do corpus compilado para esta dissertação não se debruçará sobre a ortografia, nomeadamente identificação e análise de erros

ortográficos. Desta forma, a preservação dos dados linguísticos recolhidos permitirá que o corpus mantenha a sua estrutura original e permitirá, ainda, que se consiga percecionar, de forma mais autêntica, o processo de aquisição do PE e do AP pelos informantes.

2.5. METODOLOGIA E RECOLHA DE DADOS

2.5.1. Objetivos e recolha de dados

O corpus que será objeto de estudo desta dissertação constitui dados que foram recolhidos no âmbito do projeto de investigação *Competência bilingue de crianças lusodescendentes residentes na Suíça* (Flores, Gonçalves, et al., 2022), coordenado pela Professora Doutora Cristina Flores, investigadora no Centro de Estudos Humanísticos da Universidade do Minho[35]. O referido projeto centrou-se em crianças lusodescendentes, residentes em vários cantões da Suíça (alemão, francês e italiano), com o objetivo de escrever o desenvolvimento da linguagem em falantes bilingues em contextos de migração em que o português é língua minoritária. A recolha dos dados decorreu através da aplicação de dois tipos de instrumento – (1) a elaboração de narrativas escritas e (2) a realização de um *cloze-test* (preenchimento de lacunas). Para além disso, foi distribuído um questionário de perfil linguístico, preenchido pelo encarregado de educação, a fim de recolher informações extralinguísticas sobre o tipo e a quantidade de contacto que os informantes têm com a língua portuguesa (Flores et al., 2022, p. 103). Os dados foram recolhidos entre março de 2019 e março de 2020.

2.5.2. Descrição do corpus em análise

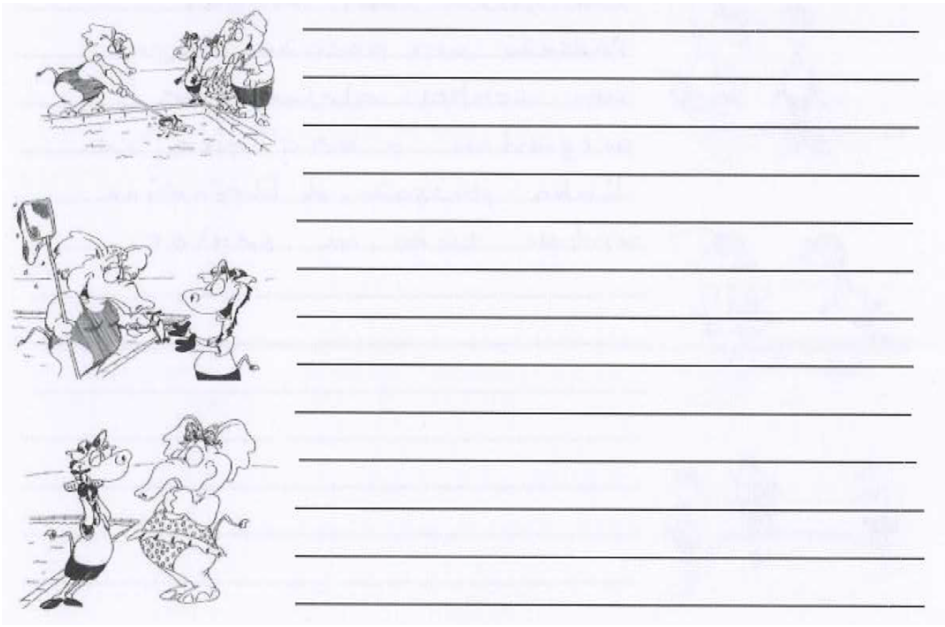
A recolha dos dados para o projeto supramencionado abrangeu cerca de 500 crianças, bem como os seus encarregados de educação (Flores et al., 2022, p. 107). Para a presente análise, foram tidos em conta apenas os dados de 20 crianças, que correspondem a 40 narrativas, dado que cada criança produziu duas narrativas. Assim, o corpus em análise nesta dissertação é constituído por 20 narrativas escritas em PE e 20 em AP. O processo de seleção das 40 narrativas que constituem o corpus em análise foi aleatório.

[35] O projeto teve parecer favorável pela Comissão de Ética para a Investigação nas Ciências Sociais e Humanas da Universidade do Minho (CEICSH 016/2019).

Para a produção das narrativas, procedeu-se à apresentação de uma pequena história infantil em português com recurso a PowerPoint (dotado de imagens e sons)[36]. Após a apresentação da história, foi entregue às crianças uma folha com algumas imagens ilustrativas referentes à história e com algumas linhas em frente às imagens para que estas pudessem recontar a história pelas suas próprias palavras (ilustração 1). Este exercício teve a duração de 15 a 20 minutos. A repetição desta tarefa deu-se com um intervalo mínimo de 2 semanas, nas mesmas condições (em contexto de sala de aula), sendo que a mesma história foi apresentada em alemão e os informantes tiveram de a reescrever, também, dentro do tempo estipulado e com recurso à mesma folha com as imagens.

Ilustração 1

Exemplo da sequência de imagens utilizada para a produção das narrativas em PE e AP.



Depois de recolhidos, os originais manuscritos foram digitalizados, transcritos para um ficheiro Excel e posteriormente divididos em orações.

[36] A apresentação do projeto de investigação e dos procedimentos adotados para a recolha dos dados foram abordados no âmbito da UC "Bilinguismo e Aquisição L2" do mestrado em Estudos Luso-Alemães (ano letivo 2019/2020) para a realização do projeto final dessa UC.

2.5.3. Perfil dos informantes

As crianças em análise nesta dissertação, têm idades compreendidas entre os 8 e os 15 anos [média = 12.07 anos], residem no cantão alemão da Suíça, sendo, por isso, falantes de português e de alemão (padrão e variante suíça). Estes alunos encontravam-se integrados na Rede de Ensino de Português do Instituto Camões, frequentando aulas de Português Língua de Herança uma vez por semana. Por sua vez, constituem um grupo particular de falantes, uma vez que adquiriram o PE e o AP como línguas maternas, em contexto migratório, sendo, por isso designados de falantes de herança.

2.7. A NARRATIVA COMO INSTRUMENTO DE INVESTIGAÇÃO

Muitos dos corpora disponíveis no repositório CHILDES visam a produção de narrativas por crianças bilingues. Roch *et al.* (2016) definem a competência narrativa como "the ability to comprehend and produce narratives, is a complex ability that involves the encoding and interpretation of information and organization of this information in a coherent mental representation" (p. 49). Por conseguinte, as narrativas produzidas por crianças refletem uma combinação de conhecimentos linguísticos e conceptuais (Tsimpli et al., 2016, p. 196). De acordo com Gagarina *et al.* (2016), a produção de narrativas é um bom instrumento para extrair dados e examinar as competências linguísticas em crianças bilingues, que pode ser feita de várias formas (p. 11). Para além de serem uma forma de comunicação presente na vida das crianças desde tenra idade e de estarem relacionada com o desenvolvimento da literacia e sucesso escolar, as narrativas são uma forma eficiente de obter informações não só sobre a estrutura da própria história, mas também sobre a linguagem da criança (Roch et al., 2016, pp. 49-50).

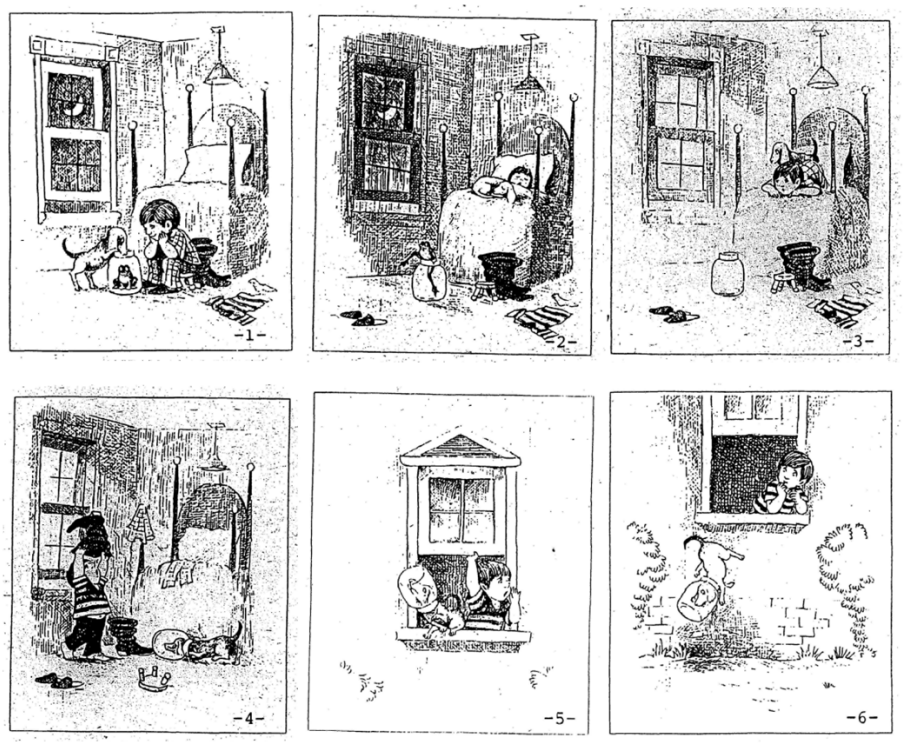
Estas, para além adquirirem um formato oral ou escrito, enquadram-se geralmente em duas categorias: pessoal e ficcional, o que geralmente ocorre tanto em contextos espontâneos como em contextos de configurações da fala criadas para o efeito (Zen, 2020, p. 92). Quanto à sua tipologia, podem ser recolhidas através de *story telling*, recontar histórias ou contar uma história após ter ouvido uma história modelo (Gagarina et al., 2016, p. 12). Assim, muitos dos corpora que se encontram no repositório CHILDES baseiam a sua produção de narrativas na "*frog story*"[37].

[37] Em 1994, Berman e Slobin publicaram estudos sobre a produção de narrativas por crianças e adultos extraídas de um livro ilustrado do escritor estadunidense Mercer Mayer intitulado *Frog, where are you?* (Bennett-Kastor, 2002, p. 131). Desde então, as "histórias do sapo" (como passaram a ser conhecidas) serviram de base para muitos outros projetos de investigação sobre a aquisição de línguas baseados na produção de narrativas (Bennett-Kastor, 2002, p. 131).

A ilustração 2 mostra um exemplo de uma sequência de imagens retiradas do livro da história do sapo.

Ilustração 2

Sequência de imagens retiradas da história "Frog, where are you?" de Marcer Mayer.



De acordo com Fiestas e Peña (2004), "the limited studies of bilingual children suggest that they may produce different narratives in each of their two languages" (p. 155). Não obstante, as autoras referem que a investigação na área de produção de narrativas por crianças bilingues não é conclusiva quanto ao facto de estas diferenças serem uma questão de variação relacionada com a proficiência linguística bilingue, diferenças estruturais linguísticas e/ou diferenças culturais relacionadas com a aquisição de cada uma das duas línguas (Fiestas & Peña, 2004, p. 155). O facto é que existem indícios que revelam uma variação ou distinção no conteúdo, estilo ou estrutura das narrativas produzidas por informantes com este tipo de perfil nas suas diferentes línguas. Muitos dos estudos realizados sobre a produção de narrativas por crianças bilingues têm sido conduzidos, principalmente, nos EUA (inglês-espanhol) (Kapalková et al., 2016, p. 146), muito provavelmente devido à expressiva comunidade de hispanofalantes que aí residem. Com

efeito, a utilização de narrativas como instrumento de avaliação e investigação da proficiência linguística em crianças bilingues destaca-se por possibilitar a análise de vários níveis linguísticos numa única tarefa, nomeadamente ao nível microestrutural (por exemplo, análise da morfossintaxe ou léxico), bem como ao nível macroestrutural, ou seja, a organização global da narração da história (Kapalková et al., 2016, p. 146). Na presente dissertação, o corpus será apenas processado através do uso de técnicas de PLN, não obstante, importa mencionar a potencialidade deste recurso como instrumento de avaliação e a sua mais-valia no âmbito do bilinguismo. Estas narrativas não só constituem uma fonte rica de dados linguísticos, como também oferecem uma visão do processo de desenvolvimento da linguagem bilingue.

No terceiro capítulo, o foco passará para uma avaliação quantitativa, debruçando-se sobre a dimensão do corpus em análise nesta dissertação e a aplicação de técnicas de PLN. Com efeito, pretende-se revelar padrões, nuances e características linguísticas significativas que podem melhorar a compreensão da aquisição de línguas em crianças bilingues.

CAPÍTULO III – ANÁLISE E REFLEXÃO SOBRE O TAMANHO DO CORPUS E O PROCESSO DE TOKENIZAÇÃO

3.1. A DIMENSÃO DO CORPUS

Dos dados recolhidos no âmbito do projeto *Competência bilingue de crianças lusodescendentes residentes na Suíça*, o corpus que se propõe como objeto de estudo desta dissertação é constituído pelos dados de 20 crianças, sendo que esses correspondem a 40 narrativas produzidas pelas mesmas (20 em PE e 20 em AP).

O tamanho do corpus é, por conseguinte, um dos principais aspetos que se tem em consideração quando se trabalha com linguística de corpus. Perguntas como: *quão grande é/será o corpus?* ou, *quantas palavras tem/terá o corpus?* são importantes desde a criação à análise do mesmo.

Como em muitas áreas de estudo, a tecnologia informática e os seus avanços foram fundamentais para que estas se pudessem desenvolver. De igual forma, estes avanços permitiram, em especial a partir dos anos 60, que o armazenamento e acesso a corpora de dimensão cada vez maior fosse viável (Hunston, 2002, p. 25). Vários autores (Fromm, 2003; Sampson & McCarthy, 2004; Sinclair, 1991; Zipf, 1935) sugerem, assim, que quanto maior for o corpus, melhor, pois mais representativo se torna. Também em muitos estudos científicos, quanto maior for a amostra mais credíveis serão os resultados. Naturalmente, a dimensão do corpus necessária para explorar questões de investigação específicas depende da frequência e da distribuição das características linguísticas (Schembri & Cormier, 2022, p. 196). Segundo Schembri e Cormier (2022), corpora utilizados para análise de questões de investigação mais específicas, por exemplo, o estudo de propriedades semânticas e sintáticas de determinados itens lexicais ou construções gramaticais menos frequentes, necessitam ter um tamanho considerável para encontrar exemplos estatisticamente significativos (pp. 196-197). Pese embora a variedade de fatores, não existe um consenso quanto ao tamanho mínimo ou máximo aceite para a criação de um corpus, sendo que este dependerá dos objetivos de pesquisa e do seu tipo (Sarmento, 2010, p. 90).

Teoricamente, o tamanho de um corpus refere-se à soma total dos seus componentes, tais como caracteres, palavras, frases, orações, entre outros (Dash & Arulmozi, 2018, p. 12). Não obstante, Sardinha (2000) propõe uma classificação dos corpora segundo o número de palavras que albergam (p. 347). Atente-se na tabela 1.

Tabela 1

Classificação de corpora quanto ao seu tamanho (Sardinha, 2000).

N.º de palavras	Classificação
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio grande
10 milhões ou mais	Grande

Na primeira coluna encontram-se o número de palavras que um corpus poderá ter e na segunda coluna consta a respetiva classificação do tamanho do corpus. Por exemplo, um corpus com menos de 80 mil palavras é considerado pelo autor, um corpus pequeno quanto à sua dimensão.

Embora a dimensão seja claramente uma vantagem importante em termos de representatividade dos dados e de generalização dos resultados, os corpora pequenos também têm um valor considerável (Granger, 2008, p. 2). Como salienta Ragan (2001), "the size of the sample is less important than the preparation and tailoring of the language product and its subsequent corpus application to draw attention to an individual or group profile of learner language use" (p. 211). Desta forma, um estudo longitudinal que tenha como objeto de estudo um único informante não perde valor pelo tamanho da sua amostra, em particular se o foco estiver na investigação do desenvolvimento individual interlinguístico (Granger, 2008, p. 3). Ademais, acrescenta Granger (2008), a dimensão de um corpus só é realmente útil se este tiver sido recolhido com base em critérios de conceção rigorosos (p. 3). Também outros autores argumentam que os corpora de pequena dimensão são úteis para estudos no âmbito da LC, como McEnery & Wilson (2001), que indicam exemplos de estudos sobre o discurso crítico em que foram utilizados corpora de pequena dimensão. Por seu turno, Anthony (2009) reforça a importância dos estudos com base em corpora de pequenas dimensões servindo-se da astronomia como analogia (p. 92). O autor refere que nesta ciência natural, alguns investigadores podem estar interessados em estudar galáxias e, a partir dessas análises, criar modelos do universo e de como este nasceu (Anthony, 2009, p. 92). Por outro lado, também existem investigadores que podem estar interessados em estudar uma única estrela, como o sol, e, assim, compreender o seu ciclo de vida, as épocas de erupção solar e padrões de emissão de radiação (Anthony, 2009, p. 92). Desta forma conclui-se que o valor de um corpus não depende do seu tamanho, mas do tipo de informação que se pode extrair do mesmo.

Independentemente do número de frases ou textos, o tamanho do corpus mede-se através do número de *tokens* que o compõe (Dash & Arulmozi, 2018, p. 10). Este é um termo basilar quando se trabalha com corpora no âmbito de estudos linguísticos, e que será alvo de uma explicação mais detalhada.

3.2. O PROCESSO DE TOKENIZAÇÃO

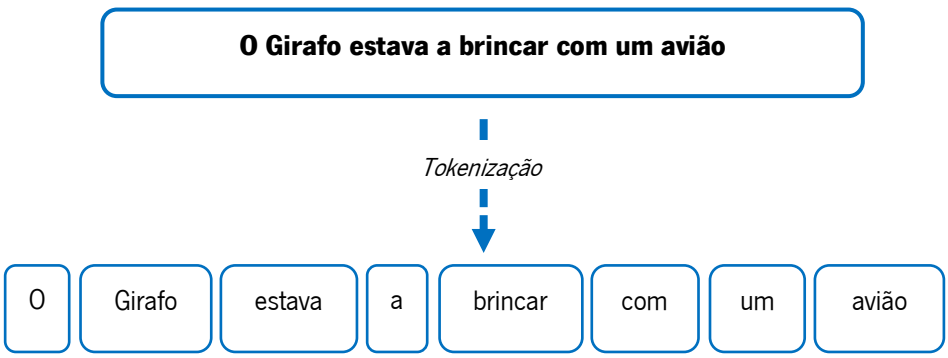
O processo de tokenização (do inglês *tokenization*) ou itemização "consiste na separação das unidades ortográficas, normalmente por meio da inserção de espaços em branco ou quebras de linha entre elas" (Sardinha, 2004, p. 128). Este processo constitui a etapa preliminar às diferentes fases de anotação de corpora (Gries, 2009; Gries & Berez, 2017), que visa a separação e/ou remoção de elementos desnecessários para o entendimento do texto e extração de conhecimento (Wynne, 2004, p. 88). Segundo Rodrigues e Teixeira (2015), a tokenização é um processo bastante simples para as línguas que usam espaços entre as palavras, como a maioria das línguas que usam o alfabeto latino (p. 15). Os autores acrescentam:

> Tokenizers often rely on simple heuristics as (1) all contiguous strings of alphabetic characters are part of one token, the same applies to numbers, and (2) tokens are separated by whitespace characters—space and line break—or by punctuation characters that are not included in abbreviations. (Rodrigues & Teixeira, 2015, p. 15)

Assim, através da tokenização obtém-se a separação do texto nos chamados *tokens* ou unidades linguisticamente significativas, como se pode observar pela figura 1.

Figura 1

Tokenização de uma oração retirada do subcorpus em PE para a exemplificação do processo.



A figura 1 constitui um esquema meramente representativo para ilustrar, de uma forma gráfica e mais simples, o desencadeamento do processo de tokenização.

No entanto, o que define em que consiste exatamente um *token* ainda é motivo de discussão. Tello (2021) lança a problemática da pontuação, que é comummente matéria de divisão entre investigadores. Por um lado, muitos consideram a pontuação como um *token*, enquanto outros defendem uma posição contrária. O autor apresenta, assim, três maneiras de como proceder com a pontuação aquando da tokenização de um texto: *(1)* pode manter-se e classificar-se como um único *token*, *(2)* pode simplesmente remover-se do texto ou *(3)* pode manter-se e juntar-se com o *token* lexical anterior ou seguinte (Tello, 2021, p. 184). Sobre que opção o autor considera mais viável, Tello (2021) explica que é preferível manter os sinais de pontuação, visto que muitos deles (como os pontos de exclamação) conferem características inerentes ao tipo e género de texto que está sob análise e contribuem para a sua diversidade (p. 184).

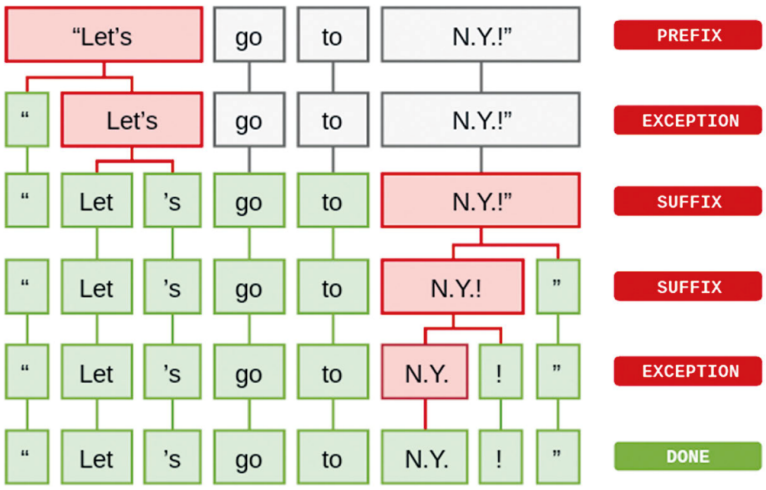
Para levar a cabo o processo de tokenização de uma maneira mais rápida e eficiente, procede-se à utilização de ferramentas que segmentem o corpus através de um tokenizador (do inglês '*tokenizer*'). Jentsch e Porada (2020) sugerem a seguinte frase *Let's go to N.Y.!* para explicar como se desencadeia este processo (p. 177). Para desempenhar esta tarefa utilizaram a ferramenta *spaCy*[38], uma biblioteca livre e *open-source* para o PLN com Python. Os autores começam por explicar que o processo de tokenização não se limita à divisão do texto por espaços em branco ou caracteres não alfanuméricos (Jentsch & Porada, 2020, p. 117). Esta tarefa é bem mais complexa. No exemplo usado por ambos, o sistema de tokenização precisa de reconhecer, por exemplo, que os pontos entre e depois das letras "N" e "Y" não indicam o fim de uma frase ou que a palavra "Let's" não constitui apenas um *token*, mas sim dois: "Let" e "us" (Jentsch & Porada, 2020, p. 117). Jentsch e Porada (2020) referem, ainda, que o processo de tokenização pode diferir, dependendo do software utilizado[39] (p. 117). A figura 2 ilustra o processo de tokenização de acordo com o exemplo supramencionado.

[38] https://spacy.io/

[39] Nem todas as ferramentas de tokenização seguem os mesmos critérios (Alyafeai et al., 2023, p. 2918). Com o tokenizador *Whitespace*, por exemplo, os sinais de pontuação são agregados ao *token* anterior, enquanto com o *WordPunctTokenizer* e o *NLTK Tokenizer* estes são processados como *tokens* únicos. Por outro lado, tokenizadores como *SentencePiece* e *BPE Tokenizer* têm a capacidade de processar *Out-of-Vocabulary* (OOV), dividindo-as em unidades menores de sub-palavras (Alyafeai et al., 2023, p. 2912). Estas são palavras que não foram previamente encontradas ou reconhecidas durante o treinamento de um modelo ou a construção de um dicionário, e, portanto, não têm uma representação direta para fins de análise ou processamento (Alyafeai et al., 2023, p. 2912).

Figura 2

Processo de tokenização utilizando o tokenizador spaCy. Figura retirada do artigo de Jentsch e Porada (2020).



Como mostra a figura 2, o tokenizador divide a frase em *tokens* através da criação de espaços em branco. Depois, inicia-se um processo iterativo no qual esta ferramenta percorre os *tokens* obtidos até que o processo não seja interrompido por mais exceções à regra/princípios de tokenização (Jentsch & Porada, 2020, p. 118), tendo, neste caso, sido obtido um total de 8 *tokens*. Uma exceção ocorre quando existem regras de tokenização que podem ser aplicadas a um *token* (Jentsch & Porada, 2020, p. 118). As regras de tokenização que desencadeiam as exceções são normalmente bastante específicas da língua utilizada no texto (Jentsch & Porada, 2020, p. 118), havendo situações particulares que não podem ser resolvidas apenas dividindo os *tokens* por caracteres de espaço, requerendo, portanto, um tratamento especial (Michelfeit et al., 2014, p. 71). No corpus em análise nesta dissertação, a pontuação assume um papel importante a ter em consideração nos princípios de tokenização, uma vez que existem marcas do discurso direto no corpus, como é possível visualizar através do pequeno excerto representado na figura 3.

Figura 3

Excerto retirado do subcorpus em PE que contém marcas do discurso direto.

Ele simplesmente tirou o avião das mãos de Girafo
e começou a brincar com o avião.
Ele ralhou com ela
a dizer: "Porque que fizeste isso?",
"Da-me o avião!" e etc.

Caso a pontuação não fosse tida em consideração no processo de tokenização, a estrutura da frase tornar-se-ia menos clara, dando origem a problemas na identificação exata das marcas do discurso direto presentes no corpus. Com efeito, a ausência da pontuação, neste caso, causaria ambiguidade na interpretação do texto. Também a remoção dos sinais interrogação e exclamação poderia resultar na perda de informação importante e levar a interpretações erróneas do corpus, como mencionado anteriormente por Tello (2021). Assim, manter os sinais de pontuação no processo de tokenização contribui a que o significado e o contexto pretendidos do texto não se alterem, sendo que estes cumprem, muitas vezes, funções sintáticas e semânticas nas frases.

Associado ao processo de tokenização está o processo de classificação dos *tokens* em *types*. A fim de dar uma explicação mais clara, Sampson e McCarthy (2004), utilizam a frase "*a rose is a rose is a rose*" para explicar que esta é composta por 8 *tokens* e 3 *types*, sendo estes *a*, *rose* e *is* (p. 3). Em termos mais simples, o número de *tokens* obtém-se pela contagem de cada palavra individual de um corpus, incluindo palavras repetidas, pontuação e números. Por oposição, os *types* correspondem apenas a uma instância única de um *token*, representando as diferentes palavras, ignorando repetições no corpus (Sarmento, 2010, p. 94). Segundo Sarmento (2010), esta relação *type-token* é normalmente obtida através da criação de listas de frequências que disponibilizam a ocorrência dos *tokens* e *types* num corpus, podendo estes estar organizados alfabeticamente e por ordem de frequência, com as formas mais frequentes em primeiro lugar (p. 94). Na tabela 2, pode observar-se uma lista de frequência com os 39 *tokens* mais frequentes do subcorpus em PE, e assim, obter uma visão geral das palavras mais frequentemente usadas. Desta forma pode identificar-se rapidamente quais são as palavras mais utilizadas no corpus (coluna *word*) e quantas vezes ocorrem no texto (coluna *frequency*). Na lista de frequência encontram-se substantivos, verbos, pronomes, entre outras partes do discurso. Através de uma lista de frequência, é, ainda, possível determinar o tamanho de um corpus, a diversidade lexical, entre outros aspetos relevantes.

Tabela 2

Lista de frequência do subcorpus em PE obtido através da ferramenta Sketch Engine.

	Word	Frequency		Word	Frequency		Word	Frequency
1	o	286	14	ele	48	27	ao	26
2	.	268	15	brincar	46	28	depois	26
3	a	213	16	uma	42	29	tentou	24
4	e	138	17	um	42	30	"	24
5	avião	134	18	muito	42	31	na	21
6	elefantina	121	19	ela	35	32	rede	21
7	girafo	96	20	estava	34	33	feliz	21
8	com	89	21	para	32	34	tinha	20
9	que	80	22	seu	29	35	brinquedo	20
10	,	77	23	da	28	36	certo	20
11	ficou	63	24	não	26	37	se	20
12	elefante	62	25	mas	26	38	dia	20
13	de	56	26	começou	26	39	triste	20

De acordo com a tabela 2, o *token* "o" ocorre com mais frequência no subcorpus em PE, demonstrando uma frequência de 286. De igual forma, *tokens* como "a", "e", "com" e "que" encontram-se entre os dez *tokens* com mais frequência na tabela 2. A este conjunto de *tokens* denomina-se de *stopwords*[40], que na sua grande maioria compreendem artigos, preposições, conjunções, entre outros elementos linguísticos. As *stopwords* são, compreensivelmente, bastante comuns em qualquer tipo de texto ou corpus (Duruttya, 2022, p. 129) e, por conseguinte, ocupam os primeiros lugares na tabela de *tokens* mais frequentes. Por outro lado, é, também, visível uma grande quantidade de substantivos e verbos, como "avião", "elefante" ou "brincar", que vão dando pistas sobre o enredo do conto usado para a produção das narrativas.

Para levar a cabo o processo de tokenização, como já foi mencionado, recorre-se a um tokenizador. Este tipo de ferramenta de PLN permite a realização desta técnica de pré-processamento do corpus. Segundo McEnery *et al.* (2019), desde 1998, têm-se registado progressos substanciais no domínio da anotação de corpus (p. 82). No seu artigo, Gamallo e Garcia (2017) apresentam algumas das opções mais conhecidas e utilizadas *suites* em PLN (p. 20), que, no entanto, são mais complexas e requerem conhecimentos no âmbito da informática,

[40] Segundo Dunn (2022), as *stopwords* (palavras vazias) são palavras que não contribuem para o entendimento de conteúdo, podendo ser ignoradas sem sacrificar o significado global de uma frase ou texto. Estas, por sua vez, são comummente removidas antes ou após o processamento de um texto em linguagem natural.

incluindo o domínio da linha de comandos. Em projetos de investigação que exigem a criação de corpora de grande dimensão, é comum que a anotação seja implementada com recurso a esse tipo de *softwares* personalizados, sendo, por sua vez, integrada em *pipelines* de ferramentas de PLN. Para tal, existem bibliotecas de *software* de código disponíveis em várias linguagens de programação, como as referidas por Gamallo e Garcia (2017), que possibilitam a anotação dos corpora. Contudo, as técnicas computacionais alargaram consideravelmente as possibilidades de utilização de um conjunto de dados linguísticos, levando à criação de ferramentas *user-friendly* para os linguistas e outros utilizadores que não estejam familiarizados com este tipo de conhecimento (McEnery et al., 2019, p. 82).

3.2.1. Ferramentas para tokenização de corpora

No seu artigo, Anthony (2013) sugere uma panóplia de ferramentas que permitem a gestão dos corpora de uma forma mais intuitiva e prática. Assim, as ferramentas gráficas (do inglês *graphical tools*) "can be useful for exploring corpus annotation, particularly when creating smaller or specialized corpora for languages with existing annotation conventions" (Newman & Cox, 2020, p. 37). Com efeito, através deste tipo de ferramentas, é possível processar corpora, através da sua tokenização, anotação morfossintática, lematização, entre outras técnicas de PLN. Das várias opções apresentadas, Anthony (2013) aconselha o manuseamento das ferramentas de "terceira e quarta gerações", por terem sido desenvolvidas mais recentemente, conterem mais funcionalidades de análise e gestão de corpora e por serem versões melhoradas de ferramentas anteriormente criadas (p. 152). Destacam-se, portanto, da terceira geração, as ferramentas *WordSmith Tools* (Scott, 1996), *MonoConc Pro* (Barlow, 2000) e *AntConc* (Anthony, 2004). De acordo com Anthony (2013), um dos grandes obstáculos das ferramentas de corpus de terceira geração em comparação com as ferramentas de quarta geração, prende-se com o facto de estas não conseguirem suportar corpora de grandes dimensões, com mais de 100 milhões de palavras (p. 152). Ademais, o autor acrescenta, "an increasing number of corpora are being released that are automatically compiled by scraping data from Internet sites" (Anthony, 2013, p. 152). Por conseguinte, corpora compilados através da recolha de dados provenientes de sítios da Internet podem albergar vários milhares de milhões de palavras e a arquitetura das ferramentas de terceira geração não se adequa para os processar (Anthony, 2013, p. 152). Tendo em conta o corpus em análise nesta dissertação, o tamanho não seria um obstáculo, visto que é consideravelmente

reduzido. Segundo o autor, outra limitação das ferramentas de terceira geração "is that publishers are becoming increasingly sensitive about allowing their data to be used for research purposes", levando a que "collections of texts can no longer be compiled and distributed for analysis with corpus tools on a personal computer" (Anthony, 2013, p. 152). Em resposta a estas problemáticas, o autor sugere no seu artigo a utilização de ferramentas de quarta geração, tais como *corpus.byu.edu* (Davies, 2013), *CQPweb* (Hardie, 2012), *Sketch Engine* (Kilgarriff et al., 2014) e *Wmatrix* (Rayson, 2008) (Anthony, 2013, p. 152). Estas, por seu turno, "offer better scalability by storing the corpus in a Web server database and pre indexing the data to allow for fast searches. They also offer protection from copyright issues by preventing users from viewing the complete corpus" (Anthony, 2013, p. 152). Anthony (2013) continua:

> Rather, users must access the corpus through a user-interface that presents only a small frame of the corpus data at a single time. The interface does, however, usually allow users to search the entire corpus and generate standard results from the entire corpus, such as KWIC concordance lines and word frequency lists. (p. 152-153)

Também através da página *Tools for Corpus Linguistics*[41] é possível aceder a uma lista bastante variada que compila 275 pacotes de software, utilizados na criação e análise de corpora. Inclui, também, informações sobre preços e sistemas operativos que suportam.

3.3. DIMENSÃO DO CORPUS EM ANÁLISE

Para averiguar o tamanho do corpus sob análise utilizou-se a ferramenta de gestão e análise de corpora *Sketch Engine*[42] (Kilgarriff et al., 2014).

As narrativas mantiveram o seu formato original, uma vez que todas as palavras foram transcritas tal e qual como estavam, mesmo contendo erros ortográfico. Segundo Chahine e Uetova (2023), "spelling is a part of linguistic command, and studies on native L1 spelling occupy an important place in pedagogical research" (p. 1). Sublinhando o papel integral da ortografia no domínio linguístico, entende-se a sua importância como uma componente-chave do domínio da língua, que, segundo as autoras, carece de atenção. Desta forma, apesar da vasta bibliografia dedicada à aquisição de uma língua estrangeira, poucos estudos se concentram na aquisição da ortografia (Chahine & Uetova, 2023, p. 1). Chahine e Uetova (2023) sublinham que "beyond its semiotic

[41] https://corpus-analysis.com/
[42] https://www.sketchengine.eu/

aspect, spelling itself represents a valuable material in understanding acquisitional processes of foreign languages. And in this respect, learner corpora open new perspectives for spelling studies" (p. 1). Apesar da análise ortográfica não constituir parte da investigação nesta dissertação, pretende-se reforçar a importância da mesma como um recurso valioso para a compreensão da forma como os indivíduos aprendem e adquirem proficiência noutras línguas que não a sua língua materna. Por conseguinte, as formas originais ortográficas das narrativas produzidas pelos informantes foram preservadas, como é possível observar pelo excerto representado na figura 4.

Figura 4

Excerto de narrativa retirada do subcorpus em PE, onde constam erros ortográficos que foram mantidos no seu processamento.

Certo dia, Giraffa e o Elefante incotraros se numa pichina.
Giraffa tinha um aveiao.
Elefante tirou a aveiao
e princou qum elle

De acordo com Rogošić (2019), "although the concept of errors refers to the violation of the rules and/or norms of the target language, the view on errors should nonetheless be a positive one as errors reflect the fact that learning is taking place" (p. 83). Desta forma, os erros são uma parte inevitável e necessária do processo natural de aprendizagem, em particular quando produzidos por falantes de herança. Ademais, para além de serem uma fonte de informação inestimável, estes revelam, muitas vezes, informações valiosas sobre os mecanismos e os desafios envolvidos no processo de aprendizagem e de aquisição de uma segunda língua, que possibilitará ao investigador uma melhor compreensão do mesmo (Afiah, 2020, p. 38).

Por sua vez, o ficheiro Excel com as narrativas foi transformado em formato .xml de forma a ser legível pelo *software* escolhido para se proceder ao processamento do corpus.

O processo de tokenização dividiu-se em duas partes: numa primeira instância procedeu-se à tokenização de um corpus geral composto por 118 narrativas produzidas por 59 informantes e numa segunda parte, a tokenização do corpus em análise, que corresponde às 40 narrativas das 118 recolhidas, produzidas por 20 dos 59 participantes. A realização deste exercício de pré-processamento do texto constitui uma etapa fundamental do processamento do corpus e permitirá a aplicação com maior precisão das tarefas subsequentes, como a etiquetação morfossintática e

a lematização do corpus (Gamallo & Garcia, 2017, p. 22). As tabelas 3 e 4 mostram os resultados obtidos em ambos os processos de tokenização.

Tabela 3

Resultados do processo de tokenização do corpus geral.

Corpus geral			
Narrativas em PE		Narrativas em AP	
Tokens	*Types*	*Tokens*	*Types*
9 355	1 478	9 328	1 184
Total			
Tokens		*Types*	
18 683		2 662	

Tabela 4

Resultados do processo de tokenização do corpus em estudo.

Corpus em estudo			
Narrativas em PE		Narrativas em AP	
Tokens	*Types*	*Tokens*	*Types*
3 780	614	3 698	603
Total			
Tokens		*Types*	
7 478		1 217	

Apesar de ser considerado um corpus pequeno, tendo em conta a divisão e classificação de Sardinha (2000) sobre o tamanho dos corpora, o corpus geral é composto por um total 18 683 *tokens* e 2 662 *types* (tabela 3). Por seu turno, o corpus em estudo nesta dissertação contém 7 478 *tokens* e 1 217 *types* (tabela 4).

3.3.1. Análise do *Type-Token Ratio* do corpus em estudo

De acordo com Nasseri e Thompson (2021), a densidade e a diversidade lexicais, enquanto duas dimensões da complexidade lexical e aspetos do conhecimento lexical produtivo, continuam a ser dois dos indicadores mais fiáveis da proficiência lexical e linguística e do desenvolvimento dos falantes nas primeira e segunda línguas (p. 1). Se por um lado a densidade lexical representa "the proportion of lexical/content words to all words/tokens", a diversidade lexical é apresentada pelos

autores como "the use of a range of diverse words (also known as unique word types) to convey meaning and is regarded as an indicator and predictor of lexical proficiency and development" (Nasseri & Thompson, 2021, p. 1). Não obstante, estes dois indicadores apesar de se inter-relacionarem, podem divergir na medida em que a densidade lexical procura mostrar a densidade dos itens lexicais nas estruturas sintáticas, enquanto a diversidade lexical é representativa da variedade dos itens lexicais e gramaticais utilizados na produção linguística (Nasseri & Thompson, 2021, p. 2). Segundo Jarvis (2013), foram propostos, ao longo dos anos, diversos índices ou medidas para calcular a riqueza lexical, acabando por se destacar o *Type-Token Ratio* (TTR), proposto em 1939 por Wendell Johnson (pp. 90-91). O TTR consiste numa medida de desenvolvimento linguístico, amplamente utilizada para calcular a proporção entre *types* e *tokens* presentes numa amostra de linguagem (Read, 2000, p. 200). Esta medida constitui uma ferramenta útil de avaliação do vocabulário, permitindo a determinação da riqueza lexical presente num corpus (Kettunen, 2014, p. 223), sendo habitualmente utilizada na análise linguística, particularmente em estudos sobre o desenvolvimento da linguagem, a aprendizagem de uma segunda língua e a complexidade dos textos. O TTR resulta da divisão efetuada entre o número de palavras diferentes (*types*) e o número de palavras produzidas na amostra (*tokens*), multiplicado por 100. A multiplicação serve para transformar o resultado em percentagem. O TTR é representado pela seguinte expressão (Operstein, 2021, p. 155):

$$TTR = \frac{\text{número de } types}{\text{número de } tokens} \times 100$$

De acordo com Williamson (2009), quanto maior for o número de *types* em relação ao número de *tokens*, mais variado é o vocabulário, ou seja, maior é a variedade lexical (p. 2). Desta forma, um TTR alto indica uma maior diversidade lexical e consequentemente menor densidade lexical, enquanto um TTR baixo indica menor diversidade e maior densidade lexical (Williamson, 2009, p. 3). É, então, esperado que um falante com alto nível de proficiência possua um vasto conhecimento vocabular que o permita evitar repetições nas suas produções escritas através do uso de sinónimos e outro tipo de palavras relacionadas (Operstein, 2021, p. 200).

Seguindo a medida proposta por Johnson, procedeu-se ao cálculo do TTR dos subcorpora em PE e AP com base nos valores obtidos e referidos na tabela 4, através do tokenizador *Sketch Engine*.

$$TTR\ subcorpus\ PE = \frac{614}{3\,780} \times 100$$

O resultado do TTR do subcorpus em PE foi de ≈ 16,24%, ligeiramente inferior ao resultado obtido com o subcorpus em AP, que foi de ≈ 16,30%, distando apenas 0,06%.

$$TTR\ subcorpus\ AP = \frac{603}{3\,698} \times 100$$

Em ambos os casos, obteve-se um resultado positivo, ainda que relativamente baixo, que se antecipava, devido à diferença do número de *types* e *tokens* dos subcorpora. Os resultados, por sua vez, evidenciam algumas características do perfil sociolinguístico dos informantes. Uma vez que se trata de um corpus bilingue, cujas narrativas foram produzidas por falantes de herança, estes valores mostram que os textos contêm um conjunto razoavelmente variado de palavras (ou itens lexicais). Por conseguinte, o vocabulário utilizado na produção das narrativas é pouco diversificado, uma vez que apenas 16% (em ambos os subcorpora) das palavras são únicas (*types*), enquanto as restantes ≈ 84% das palavras são repetições. Os resultados mostram que, apesar de haver alguma diversidade lexical, existe, ainda, uma quantidade significativa de palavras repetidas no corpus, o que pode ser uma caraterística típica em narrativas escritas por falantes de herança, que denotam um texto menos rico ou variado em vocabulário. A utilização do TTR pode, ainda, ser de grande utilidade para investigadores, professores e outros membros da comunidade científica que trabalham ou investigam matérias relacionadas com falantes de herança, na medida em que esta constitui indicador do desenvolvimento e proficiência linguística deste grupo de falantes. Ademais, através destes resultados, pode-se, também, retirar algumas conclusões e obter informações sobre a complexidade do(s) texto(s): um TTR mais elevado indica, à partida, um texto mais sofisticado, enquanto um TTR mais baixo pode sugerir um texto mais simples e repetitivo. Neste caso, por se tratar deste grupo particular de informantes, com uma média de idades de aproximadamente 12,2, a história que serviu de base para a produção das narrativas foi um conto infantil. Desta forma, e como é característico deste género literário, o conto, para além de ser bastante curto, caracteriza-se pela presença de um enredo e personagens igualmente reduzidas. Não obstante, é importante sublinhar que o TTR constitui uma medida da

riqueza lexical, sendo que outros fatores, como a estrutura gramatical, a complexidade sintática e o conteúdo, também desempenham papéis cruciais na compreensão da qualidade e outras características dos textos produzidos pelos informantes. Desta forma, é essencial considerar o TTR em conjunto com outras análises linguísticas e informações contextuais a fim de obter uma compreensão mais minuciosa do uso da língua e do seu desenvolvimento por parte deste grupo de falantes.

3.4. REFLEXÃO SOBRE OS RESULTADOS DO PROCESSO DE TOKENIZAÇÃO DO CORPUS: COMPARAÇÃO ENTRE FERRAMENTAS

Como referido anteriormente, o processo de tokenização do corpus envolve a tomada de várias decisões por parte do investigador, começando pela escolha da própria ferramenta que irá desempenhar essa tarefa. Devido à existência de uma grande variedade de *softwares* que possibilitam o pré-processamento dos corpora, é de extrema importância que se conheçam as regras de tokenização que cada um utiliza, pois estas podem tratar certas características linguísticas ou regras de tokenização de forma distinta, e assim, influenciar o processamento do corpus. Como tal, a tokenização do corpus com diferentes tokenizadores pode ser um bom exercício que ajude na eleição da ferramenta a utilizar, permitindo obter, ainda, várias perspetivas e conhecimentos sobre os dados em análise.

Posto isto, procedeu-se à tokenização dos subcorpora em PE e AP utilizando distintas ferramentas de processamento de texto, nomeadamente *Sketch Engine*, *spaCy*, *AntConc*[43] e *Voyant Tools*[44]. Nas tabelas 5 e 6 constam os resultados obtidos através dos diferentes *softwares* relativamente ao número de *tokens* e *types*.

[43] https://www.laurenceanthony.net/software/antconc/
[44] https://voyant-tools.org/

Tabela 5

Resultados dos processos de tokenização do subcorpus em PE. O subcorpus foi tokenizado com recurso a diferentes ferramentas.

Subcorpus com narrativas em PE		
Ferramenta	Tokens	Types
Sketch Engine	3 780	614
spaCy	3 755	614
Voyant Tools	3 640	543
AntConc	3 440	542

Tabela 6

Resultados dos processos de tokenização do subcorpus em AP. O subcorpus foi tokenizado com recurso a diferentes ferramentas.

Subcorpus com narrativas em AP		
Ferramenta	Tokens	Types
Sketch Engine	3 698	603
spaCy	3 673	601
Voyant Tools	3 299	542
AntConc	3 296	540

O número de *tokens* em ambos os subcorpora varia substancialmente entre os tokenizadores, principalmente se se comparar o resultado obtido através do *Sketch Engine* com os resultados obtidos pelo *Voyant Tools* e o *AntConc*. Com a primeira ferramenta obteve-se um total de 3 780 *tokens* no subcorpus em PE e de 3 698 *tokens* no subcorpus em AP, enquanto o *Voyant Tools* e o *AntConc* ocupam os últimos lugares com uma diferença de quase 300 a 400 *tokens*. A disparidade do número de *tokens* pode ser fruto de estratégias diferentes para lidar com pontuação, espaços em branco ou caracteres especiais, como os hífenes, apóstrofes, entre outros (Palmer, 2010, pp. 11-13). Os tokenizadores podem tratar este tipo de caracteres como *tokens* separados ou combiná-los com palavras adjacentes, como referido por Tello (2021). Muitos destes *softwares* procedem, também, à remoção das *stopwords* (Savoy & Gaussier, 2010, p. 458). Uma vez que esta tarefa é maioritariamente desempenhada para que as máquinas processem a linguagem humana, é comum que se eliminem palavras e outro tipo de caracteres que possam criar ruído às várias ferramentas de processamento de texto. Para além disso, a remoção das *stopwords*, permite uma redução substancial do tamanho do armazenamento dos corpora em cerca de 30% a 50% (Savoy & Gaussier, 2010, p. 458). Não obstante, a presença deste e outro tipo de elementos são fundamentais para os linguistas, pois conferem propriedades únicas aos

textos. Assim, ferramentas como o *Sketch Engine*, desenvolvida por linguistas (Kilgarriff et al., 2014, p. 16), estão mais sensíveis a estas questões aquando do processamento de corpora, incluindo a pontuação e as *stopwords* no processo de tokenização. Relativamente ao número de *types*, a variação dos resultados com as diferentes ferramentas é ainda menor. No entanto, esta desigualdade sugere uma divisão das palavras de forma diferente pelas várias ferramentas. De igual forma, o não reconhecimento de determinados elementos linguísticos específicos ou contrações, afeta o número de *types* resultantes do processo de tokenização. No caso das narrativas, muitas delas contêm erros ortográficos, e como tal, no processo de tokenização de corpora, o tokenizador, embora não reconheça certas palavras, irá contabilizá-las como um único *token*, e não havendo repetições, estas aparecerão como *types*. É, portanto, crucial selecionar o tokenizador que se alinhe com os objetivos específicos da investigação e a natureza dos dados de texto que estão a ser analisados. Ademais, pode ser benéfico explorar as regras de tokenização subjacentes e as configurações de cada tokenizador para obter uma compreensão mais profunda de seus respetivos comportamentos.

CAPÍTULO IV – CARACTERIZAÇÃO DAS ETAPAS DE ANOTAÇÃO E PROCESSAMENTO DO CORPUS: METAINFORMAÇÃO, LEMATIZAÇÃO E ANOTAÇÃO MORFOSSINTÁTICA

4.1. FUNDAMENTOS EM PROCESSAMENTO DE LINGUAGEM NATURAL

Os avanços tecnológicos das últimas décadas têm permitido, dentro de um vasto leque de ínfimas possibilidades, a criação de *corpora* eletrónicos, que através de ferramentas disponíveis no meio digital, possibilitam à comunidade científica tirar partido das suas potencialidades para fins de investigação. Assim, áreas de estudo como o PLN têm-se mostrado essenciais para o desempenho de tarefas como o processamento linguístico de textos. Segundo Puerta-Díaz *et al*. (2021), os estudos realizados no âmbito do PLN intensificaram-se desde o final da década de 1940, sendo que já nos anos 50, estes procuraram aliar a Inteligência Artificial (IA) à Linguística (p. 2). O facto é que nessa época, tanto a Linguística como a IA, ganharam mais relevância, cruzando, assim, conhecimentos num campo híbrido denominado de PLN. Esta área de estudo da IA, não só dispõe de ferramentas úteis para as ciências computacionais, como também constitui uma variedade de funções aplicáveis nas ciências que estudam matérias inerentes à linguagem, entre outras áreas do conhecimento. Por se tratar de uma área muito ativa de investigação e de constante desenvolvimento, Liddy (2001) afirma que "there is not a single agreed-upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person's definition" (p. 2). Por seu turno, Joshi (1991) salienta que a investigação em PLN é altamente interdisciplinar, envolvendo conceitos do âmbito da informática, linguística, lógica e psicologia (p. 1242). Define, ainda, que o PLN é o estudo da modelação matemática e computacional de vários aspetos da linguagem e o desenvolvimento de uma vasta gama de sistemas (Joshi, 1991, p. 1242). Coughlin (1990), numa linha mais objetiva, classifica o PLN como uma área de IA que permite ao computador processar as mesmas linguagens que os humanos utilizam na sua comunicação escrita ou oral (p. 172). O autor sublinha a importância das estratégias do PLN para os linguistas, que se estão a tornar numa parte intrínseca dos sistemas "inteligentes", desenvolvidos nas humanidades, incluindo as línguas estrangeiras (Coughlin, 1990, p. 172). Não obstante estas definições, é unânime por parte dos investigadores que um dos objetivos principais do PLN se prende em realizar um processamento da linguagem pelos computadores semelhante ao que ocorre no cérebro da espécie humana (Liddy, 2001, p. 3). Liddy (2001) refere, ainda, a existência de dois grandes focos no PLN, sendo eles (1) *language processing* e (2) *language generation* (p. 3). A autora explica que a tarefa do primeiro é equivalente ao papel do leitor/ouvinte,

enquanto a tarefa do segundo refere-se à do escritor/falante (Liddy, 2001, p. 4). Por outras palavras, o processamento de linguagem refere-se à análise da linguagem com o objetivo de produzir uma representação significativa, enquanto a geração de linguagem refere-se à produção de linguagem a partir de uma representação (Liddy, 2001, pp. 3-4). Segundo Liddy (2001), "the most explanatory method for presenting what actually happens within a Natural Language Processing system is by means of the 'levels of language' approach (...) also referred to as the synchronic model of language" (Liddy, 2001, p. 6). Como referido, estes níveis de conhecimento linguístico são utilizados para que um sistema de PLN desempenhe um processamento de linguagem natural o mais aproximado possível ao de um ser humano. De acordo com Greenberg (1998), os diferentes níveis de linguagem em PLN são (pp. 402-403):

- fonológico, que envolve a interpretação e transformação de sons em documentação auditiva;
- morfológico, que envolve a análise de segmentos de palavras tais como prefixos, sufixos, infixos, radicais e partes de palavras compostas; algumas das tarefas de PLN a este nível são a **lematização** (tópico abordado no ponto 4.7.) e a *stemização*[45];
- lexical, que envolve a determinação de significados de palavras (por exemplo, sinónimos, antónimos, ou outros tipos de relações semânticas) e pode incluir a criação ou utilização de ferramentas lexicais, tais como um glossário;
- sintático, que envolve a identificação e análise de estruturas gramaticais em texto; neste nível de conhecimento pode-se proceder, por exemplo, à **etiquetação morfossintática**/anotação gramatical de palavras (tópico abordado no ponto 4.8.);
- semântico, que envolve a interpretação do significado contextual de palavras, frases, orações e outras estruturas gramaticais num texto;
- pragmático, que envolve a criação de uma base de conhecimentos para facilitar a desambiguação.

Coughlin (1990) acrescenta que já em 1990, ano da publicação do seu artigo, o PLN era amplamente utilizado em sistemas de tradução, constituía um componente importante nos editores de texto ao nível de correções gramaticais e estilo de escrita e era, também, utilizado para o desenvolvimento de *software* de Aprendizagem Assistida por Computador (p. 172).

[45] De acordo com Korenius *et al.* (2004), o processo de stemização (do inglês '*stemming*') consiste na "normalização de *tokens*" em que a ferramenta (conhecida por '*stemmer*') reduz uma palavra flexionada ao seu radical ('*stem*') (p. 625). Este processo é bastante semelhante à lematização, não obstante, através da stemização, o radical resultante deste processo nem sempre é uma palavra válida (*non-word*) (Dash, 2021, p. 171).

Para além das várias aplicações do PLN, esta área de conhecimento dispõe, ainda, de ferramentas úteis que são essenciais para trabalhar corpora linguísticos. Assim, através do seu processamento, é possível fornecer um conjunto de dados linguísticos que poderão ser extraídos de um corpus devidamente processado e posteriormente analisados.

4.1.1. Corpus não anotado *vs.* corpus anotado

No domínio da linguística e da análise computacional, o estudo da linguagem é, como referido, facilitado pela disponibilidade de corpora, que constituem a base para várias investigações. Neste contexto, a distinção entre corpus não anotado (do inglês '*unannotated corpus*' ou '*raw corpus*') e corpus anotado ('*annotated corpus*') surge como uma consideração fundamental que afeta significativamente a profundidade e a amplitude da exploração linguística. De acordo com McEnery e Wilson (2001), "corpora may exist in two forms: unannotated (i.e. in their existing raw states of plain text) or annotated (i.e. enhanced with various types of linguistic information)" (p. 32). Entenda-se, portanto, que um corpus não anotado consiste na versão original de um conjunto de textos recolhidos numa determinada língua, sem que lhes tenha sido adicionada qualquer tipo de informação a partir de fontes externas (Dash & Arulmozi, 2018, p. 73). Em contraste, o processo de anotação de um corpus permite que este seja enriquecido com "informação linguística interpretativa" (Leech, 1997; van Rooy, 2015), realçando, desta forma, características linguísticas específicas. Também na elaboração de dicionários (ou seja, na lexicografia), a anotação é aplicada como um processo no qual o lexicografo acrescenta uma pequena nota, comentário, explicação e/ou informação de apoio, com o intuito de fornecer elementos necessários à compreensão do texto (Dash, 2021, p. 1). Neste sentido, a anotação acrescenta valor ao corpus. Segundo McEnery e Wilson (2001), "a corpus, when annotated, may be considered to be a repository of linguistic information, because the information which was implicit in the plain text has been made explicit through concrete annotation" (p. 32). Um corpus anotado contém, assim, informação introduzida por anotadores humanos ou sistemas treinados para o efeito. Leech (1997) afirma que a anotação de corpora "is widely accepted as a crucial contribution to the benefit a corpus brings, since it enriches the corpus as a source of linguistic information for future research and development" (p. 2). A anotação também é usada para descrever outros tipos de informação para além dos que podem ser adicionadas a um corpus linguístico (Hunston, 2002, pp. 19-20) ou em complemento dos níveis de anotação referidos na secção 4.1. deste capítulo. Alguns exemplos são a anotação

da entoação de um corpus falado e anotação de vários meios que representam o discurso e pensamento no texto escrito (Hunston, 2002, p. 20).

Relativamente aos níveis de anotação, Leech (1993) afirma que a informação linguística de vários níveis pode ser acrescentada a um corpus, podendo esta adquirir vários graus de granularidade (p. 275). O autor serve-se da língua inglesa como exemplo para mostrar os diferentes níveis de anotação que são geralmente aplicados nos corpora. Assim, este recurso linguístico pode ser anotado aos níveis ortográfico, fonético, prosódico, gramatical, sintático, semântico, pragmático e discursivo (Leech, 1997, p. 12). Dash (2021) refere, ainda, que alguns tipos comuns de anotação em corpora incluem a anotação anafórica, etimológica, retórica, etnográfica, de entidades mencionadas[46] e de combinatórias lexicais (*tal como; através de; em suma*) (p. 12). Atente-se na tabela 7 onde constam os dois principais tipos de anotação de texto (anotação intralinguística e extralinguística) e respetivos subtipos.

Tabela 7

Principais tipos e subtipos de anotação de texto. Adaptado de Dash (2021).

ANOTAÇÃO DE TEXTO	
Anotação intralinguística	Anotação extralinguística
Ortográfica	Semântica
Prosódica	Anafórica
Gramatical	Discursiva
Entidades mencionadas	Etimológica
Várias palavras	Retórica
Sintática	Etnográfica

Tendo em conta os níveis de anotação acima apresentados, consideram-se, assim, os mais relevantes para o estudo desta dissertação o ortográfico, gramatical e sintático. Embora a anotação ao nível ortográfico fosse pertinente neste projeto de dissertação, por se tratar de um corpus de aprendizagem e estes serem bastante ricos em aspetos ortográficos, procedeu-se, apenas, à anotação do corpus nos níveis gramatical, através da lematização, e sintático, através da etiquetação morfossintática.

[46] Segundo Marrero *et al.* (2013), o termo *Named Entity Recognition*, amplamente utilizado no PLN, é uma tarefa da extração de informação que consiste em identificar e classificar alguns tipos de elementos de informação, denominados *Named Entity*. O termo foi utilizado pela primeira vez na *6th Message Understanding Conference*, onde ficou clara a importância da identificação semântica de pessoas, organizações e localizações, bem como de expressões numéricas como tempo e quantidades.

4.1.2. Informação intralinguística, extralinguística e extratextual

De acordo com Dash e Arulmozi (2018), "while an unannotated corpus contains simple raw state of plain texts, an annotated corpus contains texts that are encoded with extralinguistic and intralinguistic information of different types" (p. 73). Segundo os autores, as questões intralinguísticas referem-se a toda a informação relacionada com o conteúdo do corpus, ou seja, relacionada com a ortografia, parte do discurso, morfologia, gramática, pragmática, entre muitas outras propriedades linguísticas (Dash & Arulmozi, 2018, p. 73). Por outro lado, os elementos extralinguísticos ajudam o utilizador a conhecer a informação com base na qual se consegue ter uma melhor interpretação do texto (Dash, 2021, p. 3). Dash (2021) explica:

> For instance, when we annotate a text at the anaphora level, there is no apparent information available in the text based on which we identify how words are linked with others in a relation of coreference. We have to go beyond the textual level to understand how words are co-indexed or bound with coreferential relations. (p. 14)

Um corpus pode, ainda, conter informações provenientes de vários domínios externos. Estas informações são identificadas como informações extratextuais, que podem ser sistematicamente codificadas e preservadas num ficheiro de cabeçalho de um corpus eletrónico (Dash, 2021, p. 72). Ädel (2020) acrescenta: "markup allows the corpus builder to include important information about each file in the corpus" (p. 13). Numa primeira instância, a informação extratextual pode parecer irrelevante, não obstante, através desta é possível preservar e garantir os direitos de autor, bem como potenciar a investigação em corpora através da análise de questões relacionadas com a sociolinguística (Dash & Arulmozi, 2018, pp. 74-75). De acordo com Dash (2021), "by definition, an extratextually annotated corpus is furnished with additional extratextual annotation as metadata along with some basic information related to the text" (p. 72). Neste sentido, a informação extratextual deve ser adicionada ao corpus sob a forma de metainformação. Aludindo às múltiplas vantagens da metainformação, Lange (2022) afirma que as suas duas principais funções são permitir que os dados linguísticos sejam encontrados e também que sejam compreendidos, fornecendo um contexto adicional (p. 108). De igual forma, Ädel (2020) refere "with better metadata about individual texts and speakers, we will be in a better position to understand the data, not only to correlate metadata to variation, but also to see more precisely how corpora differ in the case of comparison" (p. 22). Com efeito, através da metainformação, a capacidade de correlacionar informações e identificar variações nos corpora é reforçada, dado que estes

desempenham um papel fundamental na facilitação de uma análise precisa dos dados linguísticos.

4.2. VANTAGENS DA ANOTAÇÃO DE CORPUS

Devido à sua vasta aplicabilidade, os corpora são utilizados, como recurso linguístico, em vários domínios. Gries e Berez (2017) começam por explicar que "the types of information corpora are annotated for is dependent on the kind, and thus typicality, of corpus, i.e. the way in which the data have been collected" (p. 382). Por outras palavras, a escolha das informações a serem anotadas num corpus está intimamente ligada ao propósito do mesmo e à maneira como os dados foram recolhidos. Os autores destacam, desta forma, que o processo de anotação não é uniforme para todos os tipos de corpora, devendo ser, por isso, adaptado para atender às necessidades e objetivos específicos de cada tipo de corpus. Hunston (2002) refere que "a corpus essentially tells us what language is like, and the main argument in favour of using a corpus is that it is a more reliable guide to language use than native speaker intuition is" (p. 20). A autora sugere que os corpora constituem uma ferramenta que permite compreender como a linguagem é utilizada, dado que oferecem uma visão objetiva e baseada em evidências da língua em contexto real. Ademais, este recurso linguístico quando anotado, pode potenciar a sua usabilidade. Dash (2021) enumera algumas áreas cujos textos das línguas naturais são utilizados para o seu sustento e crescimento, tais como na IA, Tecnologia da Informação, Tecnologia da Linguagem, Linguística Computacional, História, Estudos Culturais, Sociologia, Ciências Cognitivas e Ciências Forenses (pp. 8-9). Segundo o autor, muitas destas disciplinas utilizam textos linguísticos, tanto na sua versão não anotada como anotada, e trabalham com base nos resultados obtidos a partir dos mesmos (p. 9). Por outro lado, Hovy e Lavid (2010) apontam que "neither manual nor automated annotation is infallible, and both have advantages" (p. 14). Desta forma, e apesar do artigo de Hovy e Lavid (2010) ser bastante anterior ao livro de Dash (2021), a verdade é que a anotação de textos, seja manual ou automática, resulta de um processo falível, carecendo, por vezes, de algumas falhas. Pese embora essas limitações, vários estudos apontam que a utilização de textos anotados é uma mais-valia para o estudo e análise de fenómenos linguísticos, no processamento de textos e nas aplicações de textos em vários domínios do conhecimento humano (Dash, 2021, p. 9). Newman e Cox (2020) acrescentam "(...) annotated corpora offer great advantages over the raw text when it comes to the investigation of linguistic phenomena" (p. 25). Tendo em conta a perspetiva de

Leech (1997), um corpus anotado permite a extração de informação, reutilização e multifuncionalidade de textos (pp. 4-5). O primeiro ponto refere-se à extração de informação linguística, uma vez que o corpus pode ser etiquetado em vários níveis, enquanto o segundo ponto refere-se ao fenómeno de que um mesmo corpus anotado pode ser utilizado por vários utilizadores para vários fins de investigação (Leech, 1997, pp. 4-5). Por último, o terceiro ponto refere-se ao facto de a aplicação de um corpus anotado não se limitar a alguns domínios da linguística, como a lexicografia ou o ensino das línguas (Leech, 1997, p. 5). Pelo contrário, segundo o autor, um corpus anotado destina-se a ser utilizado em muitos domínios e disciplinas que não estão diretamente ligados à linguística como, aliás, já fora referido. Tendo em conta estes fatores, a aplicação de textos anotados pode traduzir-se em domínios como geração de textos anotados legíveis por computador, geração de recursos linguísticos digitais, ferramentas de tecnologia linguística e desenvolvimento de sistemas, ferramentas de apoio à tradução, desenvolvimento de interfaces interativas, geração de recursos personalizados para indústrias de TI, ferramentas de tecnologia da fala e por último, geração de recursos para fins académicos e comerciais (Dash, 2021, p. 9).

4.3. INICIATIVA DE CODIFICAÇÃO TEXTUAL: *Text Encoding Initiative*

Através da emergente análise sistemática dos corpora, que revolucionaram a investigação linguística, tornou-se necessária a procura de um modelo de codificação de texto eletrónico a vários níveis que possibilitasse a preservação e acessibilidade deste recurso linguístico no meio digital. Lange (2022) afirma que existem vários modelos de codificação de metainformação estabelecidos para corpora e que estes podem variar em termos de expressão e a sua utilização pode também depender do formato de ficheiro utilizado para os próprios dados do corpus (p. 108). Desta forma, quando compilados, os corpora precisam de ser codificados e anotados para serem armazenados e processados posteriormente (Piotrowski, 2012, p. 53) e nesse sentido, a *Text Encoding Initiative*[47] (TEI) surgiu como uma solução pioneira para o desafio. Segundo Lange (2022):

> Because TEI is more dominant for written data, TEI headers are more relevant for corpora containing written learner data, but it should be noted that the TEI guidelines also cover

[47] https://tei-c.org/

> transcription of spoken language which would make TEI headers also relevant for spoken learner data. (p. 109)

Até aos anos 80, cada académico, projeto, *software* ou grupo de investigação criava o seu próprio sistema de representação, característico por ser estrutural ou tipográfico (Vanhoutte, 2004, p. 10). Como Vanhoutte (2004) explica, esta miscelânea deu origem a uma série de esquemas de codificação cuja funcionalidade, na maioria das vezes, não podia transcender os limites do projeto ou da ferramenta para a qual tinha sido criada (p. 10). Assim, em apelo à então conjuntura dos factos surgiu a *Standard Generalized Markup Language* (SGML), que se tornou uma norma ISO (*International Organization for Standardization*) – SGML-ISO 8879 – em 1986 (Goldfarb, 1990, p. 8). Segundo Vanhoutte (2004), "SGML is not in itself a markup scheme, but a methodology that enables the creation of such schemes" (p. 10). Após a publicação da SGML, um grupo de 32 académicos da área das humanidades computacionais reuniram-se na *Vasser College* em Poughkeepsie (Nova Iorque) e acordaram sobre um conjunto de princípios metodológicos – os chamados *Poughkeepsie Principles*[48] – que constituíram a base da TEI (Vanhoutte, 2004, p. 10). Desenvolvida a partir de 1987 (Sperberg-McQueen, 1994, p. 409), a TEI surgiu como uma estrutura robusta, com o objetivo de fornecer diretrizes independentes para a codificação e anotação de textos de uma forma normalizada e legível por máquina (Piotrowski, 2012, p. 60). Não obstante, devido à complexidade da SGML e com os recentes avanços na codificação de texto, em particular com o desenvolvimento da *Extensible Markup Language* (XML), a TEI migrou para o uso da XML como linguagem de anotação para implementar as suas diretrizes (Schreibman, 2002, p. 284), contribuindo, assim, para uma anotação mais prática, padronizada e viável em ambientes digitais. De acordo com Piotrowski (2012):

> XML is the standard for adding higher-level information – markup and metadata – to texts (...) and [it] also plays an important role in NLP, especially for annotating corpora and texts, for example with information on tokens, sentence boundaries, part-of-speech tags, morphological analyses, chunking, or named entities. (p. 60)

Walker (1994) alerta que a TEI "is equally concerned with equations, figures, tables, spoken language, diagrams, images, and hypertext linkages" (p. 368). As diretrizes da TEI foram alvo de várias revisões, tendo a última (*P5 Guidelines*) sido lançada em 2007 (Piotrowski, 2012, p. 61). Em 2000, a TEI foi constituída como uma organização sem fins lucrativos – *the TEI Consortium*[49]

[48] https://tei-c.org/Vault/ED/edp01.htm
[49] https://tei-c.org/about/mission/

– para apoiar e incentivar o desenvolvimento desta estrutura (Vanhoutte, 2004, p. 14). As diretrizes da TEI estão publicadas sob o nome de *Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange*[50] e estabeleceram-se como uma norma internacional, amplamente utilizada por projetos de investigação e instituições nas ciências humanas para codificar todos os tipos de material textual (Piotrowski, 2012, p. 61). Dash (2021) aponta que "the application of TEI on corpora has been a huge success as it provides standards for corpus annotation as well as opens ways for building text interchanging facilities in machine-readable form" (p. 78). Através das diretrizes da TEI, investigadores, académicos, e outros utilizadores podem criar edições digitais de textos, compilar corpora e partilhar dados de forma mais eficaz. Gibson e Ruotolo (2003) sublinham, assim, três importantes características desta diretiva de etiquetação – "*stable, extensible and malleable*" (p. 57). A primeira refere-se à robusta estabilidade da TEI, cujo conteúdo anotado através das suas diretrizes não será afetado pela obsolescência de plataformas informáticas ou programas de *software* específicos; a segunda característica ressalta a capacidade da TEI em acompanhar a evolução da tecnologia e as necessidades de anotação dos utilizadores; e por último, os autores apontam a flexibilidade da TEI em permitir a reconfiguração e a reutilização de dados numa variedade de formatos e potenciais usos (Gibson & Ruotolo, 2003, p. 57). A TEI desempenha um papel crucial no domínio das humanidades digitais, permitindo a criação de textos bem estruturados, normalizados e acessíveis, garantindo, também, que os recursos textuais mantenham o seu valor académico e possam ser utilizados para uma vasta gama de fins de investigação.

No plano mais operacional, as diretrizes da TEI traduzem-se num vasto conjunto de etiquetas definidas com o objetivo de codificar informação de vários tipos nos corpora. Todos os documentos TEI contêm, no início do documento, um bloco de metainformação – The *TEI Header* (Piotrowski, 2012, p. 61). No corpus em estudo, essa secção será de extrema importância, pois aí constarão, devidamente codificadas, as informações obtidas através dos questionários sociolinguísticos. Segundo Lange (2022), o cabeçalho TEI pode conter cinco partes principais, das quais apenas a primeira é obrigatória: (1) a descrição do ficheiro (*File Description*), que contém uma descrição bibliográfica completa do próprio ficheiro informático; (2) uma descrição da codificação (*Encoding Description*), que descreve a relação entre o texto eletrónico e a sua fonte ou fontes; (3) um perfil do texto (*Text Profile*), que contém informações classificatórias e contextuais sobre o texto, tais

[50] https://tei-c.org/guidelines/

como o seu assunto, o contexto da sua criação, os participantes, entre outros aspetos; (4) um *Container Element* para a inclusão de outros formatos de metainformação diferentes da TEI e (5) um historial de revisões (*Revision History*), que fornece um historial das alterações efetuadas durante o desenvolvimento do texto eletrónico (p. 109). Como referido, a secção obrigatória *File Description* (<fileDesc>) contém, por sua vez, três subsecções obrigatórias: *Title Statement* (<titleStmt>), *Publication Statement* (<publicationStmt>) e *Source Description* (<sourceDesc>), como é possível observar através da ilustração 3. Desta forma, acrescenta Piotrowski (2012), o cabeçalho TEI permite uma exposição muito pormenorizada do corpus codificado, podendo este, exceder em tamanho o próprio corpus (p. 61). Através da ilustração 3, pode observar-se um exemplo de um cabeçalho TEI minimalista, retirado do sítio da TEI.

Ilustração 3

Exemplo de um cabeçalho TEI minimalista. Exemplo retirado do sítio da TEI.

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>
<!-- title of the resource -->
   </title>
  </titleStmt>
  <publicationStmt>
   <p>
<!-- Information about distribution of the resource -->
   </p>
  </publicationStmt>
  <sourceDesc>
   <p>
<!-- Information about source from which the resource derives -->
   </p>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

4.4. METAINFORMAÇÃO

Segundo Ädel (2020):

> Metadata can consist of different types of information. For example, the corpus compiler may include information based on interviews with participants or participant observation. A common way of collecting metadata is by asking corpus participants to fill out a questionnaire which has been carefully designed by the corpus compiler so as to include information likely to be relevant with respect to the specific context of the discourse included and the people represented. (p. 11)

No corpus em estudo, a metainformação é de extrema importância, pois através de questionários sociolinguísticos distribuídos pelos familiares dos informantes, possibilitou-se a recolha de dados que permitem o cruzamento de variáveis. Correia e Flores (2021) explicam:

> Para avaliação do efeito da experiência linguística sobre o desenvolvimento bilingue, os estudos recorrem, geralmente, a questionários sociolinguísticos, os quais permitem ao investigador traçar o perfil sociolinguístico dos sujeitos em análise, bem como obter informações cruciais sobre variáveis preditivas do desenvolvimento bilingue. (p. 76)

No seu artigo, as autoras apresentam um questionário sociolinguístico – Questionário Sociolinguístico Parental para Famílias Emigrantes Bilingues[51] (QuesFEB) –, desenvolvido com o objetivo de recolher dados sobre a experiência sociolinguística de crianças bilingues com *background* migratório (Correia & Flores, 2021, p. 76). Através desse questionário, que à data sofreu ligeiras adaptações, foi possível recolher os dados sociolinguísticos que serviram de base para o projeto de investigação Flores *et al.* (2022) e para esta dissertação. De acordo com Correia e Flores (2021):

> No caso específico de estudos centrados no desenvolvimento linguístico de crianças falantes de herança, os questionários sociolinguísticos tendem a ser aplicados, de forma direta ou indireta, ao cuidador da criança, sendo-lhe solicitado que forneça informações biográficas sobre o agregado familiar, bem como sobre a experiência sociolinguística não só da criança, mas também, de acordo com os objetivos de cada estudo, de outros indivíduos que contactem regularmente com a mesma. (p. 82)

O QuesFEB constitui, assim, um questionário parental desenvolvido a partir do trabalho de doutoramento de Liliana Correia, que se destina à comunidade científica que pretenda realizar estudos sobre a aquisição de línguas de herança por crianças bilingues (Correia & Flores, 2021, p. 87). É um questionário que pode ser preenchido autonomamente (sem ser necessária a presença do investigador), por um dos encarregados de educação ou cuidadores da criança e demora aproximadamente 20 minutos a ser preenchido (Correia & Flores, 2021, p. 88). Desta forma, o QuesFEB é composto por três secções, sendo que as duas primeiras partes se destinam à recolha de informações biográficas e sociolinguísticas dos pais da criança e a terceira parte é centrada na criança. Através da figura 5, pode observar-se um levantamento (adaptado) das informações recolhidas através do questionário sociolinguístico para este projeto de dissertação.

[51] https://doi.org/10.34622/datarepositorium/1JSLNQ

Figura 5

Levantamento dos itens do questionário sociolinguístico utilizado para a recolha de metainformação (adaptado).

Questionário sociolinguístico

PARTE I – Sobre o Encarregado de Educação

1. Idade
2. Grau de parentesco
3. País de nascimento
4. Há quantos anos vive na Suíça
5. País de nascimento dos pais
6. Região em PT da qual a família é originária
7. Profissão
8. Escolaridade
9. Domínio do PE (fala e compreensão)
10. Domínio do AP (fala e compreensão)
11. Domínio do CH (fala e compreensão)
12. Língua usada (na Suíça):
    a. Com familiares de PT que vivem na Suíça
    b. Com amigos de PT que vivem na Suíça
    c. No trabalho
    d. Na igreja
    e. No restaurante
    f. Nas compras
    g. Em clubes/associações de PT
    h. Para ver televisão
    i. Para ler livros/jornais/revistas
    j. Para ouvir música/rádio
    k. Na internet
    l. Outro
13. Importância do PE na vida do educando
14. Foi aconselhado a não falar PE com o seu educando
    a. Se sim, falou menos PE com o seu educando
15. Encoraja o seu educando a aprender/falar PE
    a. Se sim, indique as razões
    b. Se não, indique as razões para não encorajar o seu educando a aprender/falar PE

Parte II – Sobre o/a pai/mãe

1. Idade
2. Grau de parentesco
3. País de nascimento
4. Há quantos anos vive na Suíça
5. País de nascimento dos pais
6. Região em PT da qual a família é originária
7. Profissão
8. Escolaridade

Parte III – Sobre a criança (informante)

1. Nascimento
2. País de nascimento
3. Idade em que chegou à Suíça
4. Irmãos:
    a. Data e país de nascimento
    b. Idade em que chegou à Suíça
5. Problemas de saúde/desenvolvimento
    a. Audição
    b. Fala (perturbação da linguagem)
        i. Frequentou terapia da fala
6. Quando foi primeiro contacto com:
    a. PE
    b. CH
    c. AP

7. Como se deu o primeiro contacto com PE
8. Como/quando foi primeiro contacto com AP e CH
9. Outra língua que fale/ouça no dia a dia
    a. Primeiro contacto com essa língua
    b. Como se deu esse contacto
10. Proficiência em PE (compreensão e fala)
11. Proficiência em AP (compreensão e fala)
12. Proficiência em CH (compreensão e fala)
13. Agregado familiar
14. Regularidade com que fala PE e Alemão (AP ou CH) com o seu filho
15. Regularidade com que o seu filho fala PE e Alemão (AP ou CH)
16. Avós PT a viver na Suíça
17. Horas em contacto com PE e Alemão
18. Variante (AP ou CH) em que o seu filho realiza atividades
19. Tempo frequenta aulas de PE
    a. Horas na semana
20. Participa em atividades sobre língua e cultura PT
21. Vai de férias a PT
22. Como é a comunicação em PT nas férias
23. Comentários/observações

Segundo Stoll e Schikowski (2020), "any corpus relies on metadata to correlate the speech of the child and her surroundings with social variables" (p. 315). Assim, foi possível obter informação relevante sobre os familiares, os contextos e os próprios informantes, tais como a idade, o nível de proficiência, qual a língua materna, contexto de aprendizagem da língua, formação académica, tempo de exposição a uma determinada língua, entre outros aspetos. Ademais, a metainformação possibilita o cruzamento de dados, através de comparações entre diferentes grupos de informantes, métodos de ensino ou níveis de proficiência linguística, que possibilitam a identificação de padrões e diferenças na utilização e desenvolvimento de competência linguística nas línguas em estudo. Ädel (2020) refere que "various types of metadata can be placed in a separate file or in a 'header', so that a computer script or web-based tool for example will be able to use the information in systematic ways when counting frequencies, searching for or displaying relevant data." (p. 13). Ädel (2020) refere, ainda:

> In a language-learning context, some of the variables likely to be relevant include what the learner's first language is, what the medium of instruction was in school, how much exposure the learner has had to the second language – whether through instruction in a school context or through spending time in a context where the second language is spoken. (p. 11)

A autora aponta algumas variáveis de extrema importância, das quais, através do questionário sociolinguístico acima representado, se podem cruzar dados e, por exemplo, estudar potenciais diferenças entre informantes que já nasceram na Suíça ou que emigraram durante a infância; ou

até mesmo, consoante os níveis de exposição a que as crianças estão ao português, testar uma hipótese de investigação sobre a aquisição de uma segunda língua.

No seguimento da recolha dos dados sociolinguísticos, segue-se a transformação das informações contidas no modelo do questionário (figura 5), bem como as respostas obtidas, num esquema de metainformação. Com efeito, apresenta-se uma proposta de um esquema de metainformação em formato XML, recorrendo às diretrizes implementadas pela TEI (figura 6). Importa referir que o questionário em questão é extenso, com variáveis específicas ao projeto para o qual foi criado e de vária ordem. Para além disso, foi concebido para o preenchimento em papel e não para o formato eletrónico. Assim, a transposição dos resultados do questionário para formato eletrónico e a respetiva anotação apresentou desafios nomeadamente ao nível dos itens de resposta aberta. Por conseguinte, adotou-se uma abordagem híbrida: utilização das diretrizes da TEI que se adequam ao tipo de informação e propostas próprias adequadas às informações específicas do questionário.

Figura 6

Proposta de esquema de metainformação em XML com recurso às diretrizes implementadas pela TEI.

```
<teiCorpus>
   <teiHeader>
      <!–Metainformação do Corpus PE–>
   </teiHeader>
   <TEI>
      <teiHeader>
         <!–Metainformação da narrativa 1–>

         <fileDesc>
            <titleStmt>
               <title>Elefantina, Girafo e o avião: versão eletrónica</title>
            </titleStmt>
            <publicationStmt>
               <authority>
                  <name>Cristina Flores</name>
                  <name>Idalete Dias</name>
               </authority>
               <availability status="free"/>
            </publicationStmt>
            <sourceDesc>
               <p>Uma folha com imagens ilustrativas referentes à história intitulada “Elefantina, Girafo e o avião” e a
narrativa escrita por uma criança falante do Português Europeu como língua de herança.</p>
            </sourceDesc>
         </fileDesc>

         <encodingDesc>
            <projectDesc>
               <p>Narrativa recolhida no âmbito do projeto de investigação “Competência bilingue de crianças
lusodescendentes residentes na Suíça”.</p>
               <p>O projeto centrou-se em crianças lusodescendentes, residentes em vários cantões da Suíça (alemão,
```

```
francês e italiano), com o objetivo de escrever o desenvolvimento da linguagem em falantes bilingues em contextos de migração em que o português é língua minoritária</p>
            </projectDesc>
        </encodingDesc>

        <profileDesc>
            <creation>
                <date from="2019-03" to="2020-03">Março 2019 - Março 2020</date>
                <place>Suiça</place>
            </creation>
            <langUsage>
                <language ident="PE">Português Europeu </language>
            </langUsage>
            <particDesc>
                <listPerson>
                    < person xml:id="CH002_PTAL_CC_100519_M" ident="mother">
                    <age>38 anos</age>
                    <relation>Mãe</relation>
                    <country type="birth_mother">
                        <birthPlace type="mother">Portugal</birthPlace>
                    </country>
                    <residence type="duration" country="CH" value="20">Vive na Suiça há 20 anos</residence>
                    <country type="birth_grandparents">
                        <birthPlace type="grandmother">Portugal</birthPlace>
                        <birthPlace type="grandfather">Portugal</birthPlace>
                        </county>
                        <origPlace type="region_PT">Amarante</origPlace>
                        <occupation>Limpeza doméstica</occupation>
                        <education>
                            <level>1.º ciclo (1.º-4.º ano)</level>
                        </education>
                        <langKnowledge lang="PT">
                            <langKnown type="spoken" level="master">Fluente (como uma língua materna)</langKnown>
                            <langKnown type="comprehension" level="master">Perfeita (como uma língua materna)</langKnown>
                        </langKnowledge>
                        <langKnowledge lang="StandardDE">
                            <langKnown type="spoken" level="basic">Mais ou menos bem (consigo manter apenas uma
conversação simples em Alemão Padrão)</langKnown>
                            <langKnown type="comprehension" level="basic">Mais ou menos bem (consigo perceber apenas uma
conversação simples e em ritmo lento)</langKnown>
                        </langKnowledge>
                        <langKnowledge lang="StandardDE_vs_CH">
                            <langKnown type="spoken" level="+StandardDE">Melhor Alemão Padrão do que Alemão
Suíço</langKnown>
                            <langKnown type="comprehension" level="+StandardDE">Melhor Alemão Padrão do que Alemão
Suíço</langKnown>
                        </langKnowledge>
                        <listContext type="langUsage_DE_CH">
                            <context>
                                <label>Com familiares portugueses a viverem na Suíça</label>
                                <value>Apenas Português</value>
                            </context>
                            <context>
                                <label>Com amigos portugueses a viverem na Suíça</label>
                                <value>Apenas Português</value>
                            </context>
                            <context>
                                <label>No trabalho</label>
                                <value>Apenas Alemão</value>
                            </context>
                            <context>
                                <label>Na igreja</label>
                                <value>Apenas Português</value>
                            </context>
```

```
                    <context>
                        <label>No restaurante</label>
                        <value>Português e Alemão em quantidades iguais</value>
                    </context>
                    <context>
                        <label>Nas compras (padaria, supermercado, etc)</label>
                        <value>Português e Alemão em quantidades iguais</value>
                    </context>
                    <context>
                        <label>Em clubes e/ou associações portuguesas</label>
                        <value>Mais Português do que Alemão</value>
                    </context>
                    <context>
                        <label>Para ver televisão</label>
                        <value>Apenas Português</value>
                    </context>
                    <context>
                        <label>Para ler livros/revistas/jornais</label>
                        <value>Mais Português do que Alemão</value>
                    </context>
                    <context>
                        <label>Para ouvir música/rádio</label>
                        <value>SR</value>
                    </context>
                    <context>
                        <label>Para navegar na internet</label>
                        <value>Mais Português do que Alemão</value>
                    </context>
                </listContext>
                <relevance type="langknowledge" lang="PT" value="high">
                    <listReason>
                        <reason n="1">Comunicar com a família alargada (avós, tios, primos, etc);</reason>
                        <reason n="2">Ser vantajoso/bom para o sucesso escolar do(a) meu(minha) filho(a) (por exemplo,
para ele/ela ter melhores resultados na escola, etc);</reason>
                        <reason n="3">Ser vantajoso/bom para o sucesso profisisonal do(a) meu(minha) filho(a) (por
exemplo, para conseguir emprego melhor no futuro, etc);</reason>
                        <reason n="4">Regressar a Portugal no futuro;</reason>
                    </listReason>
                </relevance>
                </person>

                <person xml:id=" CH002_PTAL_CC_100519_F" ident="father">
                    <age>39 anos</age>
                    <relation>Pai</relation>
                    <country type="birth_father">
                        <birthPlace type="father">Portugal</birthPlace>
                    </country>
                    <residence type="duration" country="CH" value="18">Vive na Suiça há 18 anos</residence>
                    <country type="birth_grandparents">
                        <birthPlace type="grandmother">Portugal</birthPlace>
                        <birthPlace type="grandfather">Portugal</birthPlace>
                    </country>
                    <origPlace type="region_PT">Amarante</origPlace>
                    <occupation>Carpinteiro</occupation>
                    <education>
                        <level>3.º ciclo (7.º-9.º ano)</level>
                    </education>
                </person>

                <person xml:id="CH002_PTAL_CC_100519" ident="child">
                    <birth when="2007-06-09">09-06-2007</birth>
                    <country type="birth_child">
                        <birthPlace type="child"> Suíça</birthPlace>
                    </country>
```

```
                    <siblings>
                        <option value="no">Não</option>
                    </siblings>
                    <condition type="health">
                        <problem type="hearing">Não</problem>
                        <problem type="speech">Não</problem>
                    </condition>
                    <firstContact lang="PT">
                        <when value="birth">quando nasceu</when>
                        <contactPersons>
                            <option value="mother">Mãe</option>
                            <option value="father">Pai</option>
                            <option value="grandparents">Avós</option>
                            <option value="Uncles_Cousins">Tios e primos</option>
                        </contactPersons>
                    </firstContact>

                    <firstContact lang="DE_CH">
                        <when value="5 years" >quando tinha 5 anos</when>
                        <where value="pre-school">Pré-escola</where>
                    </firstContact>

                    <firstContact lang="StandardDE">
                        <when value="5 years">quando tinha 5 anos</when>
                        <where when="pre-school">Pré-escola</where>
                    </firstContact>
                    <firstContact lang="other">
                        <option value="no">não</option>
                    </firstContact>

                    <langKnowledge lang="PT">
                        <langKnown type="speech" level="advanced">Muita fluência/Sem nenhuma
dificuldade</langKnown>
                        <langKnown type="comprehension" level="advanced">Compreende muito bem, com quase
nenhuma dificuldade</langKnown>
                    </langKnowledge>
                    <langKnowledge lang="StandardDE">
                        <langKnown type="speech" level="intermediate">Bastante fluência/Com poucas
dificuldades</langKnown>
                        <langKnown type="comprehension" level="intermediate">Compreende bem, com poucas
dificuldades</langKnown>
                    </langKnowledge>
                    <langKnowledge type="DE_CH">
                        <langKnown type="speech" level="advanced">Muita fluência/Quase nenhumas
dificuldades</langKnown>
                        <langKnown type="comprehension" level="advanced">Compreende muito bem, com quase
nenhuma dificuldade</langKnown>
                    </langKnowledge>

                    <residence type="grandparents">
                        <country>Portugal</country>
                    </residence>

                    <listEvents type="langUsage_activities" lang="PT_vs_DE">
                        <event>
                            <label>Ver televisão</label>
                            <value>Sempre em Português, nunca em Alemão</value>
                        </event>
                        <event>
                            <label>Ouvir música</label>
                            <value>Metade em Português e metade em Alemão</value>
                        </event>
                        <event>
                            <label>Ler livros/revistas</label>
```

```
                        <value>Metade em Português e metade em Alemão</value>
                    </event>
                    <event>
                        <label>Falar com familiares/amigos de Portugal (telefone, Skype, Messenger, WhatsApp)</label>
                        <value>Sempre em Português, nunca em Alemão</value>
                    </event>
                    <event>
                        <label>Ver vídeos no YouTube</label>
                        <value>Metade em Português e metade em Alemão</value>
                    </event>
                </listEvents>
                <listEvents type="langUsage_activities" lang="StandardDE_vs_DE_CH">
                    <event>
                        <label>Ouvir música</label>
                        <value>Alemão Padrão</value>
                    </event>
                    <event>
                        <label>Ler livros/revistas</label>
                        <value>Alemão Padrão</value>
                    </event>
                    <event>
                        <label>Ver vídeos no YouTube</label>
                        <value>Alemão Padrão</value>
                    </event>
                </listEvents>
                <studyHistory lang="PT">
                    <period value="5 years">Há 5 anos</period>
                    <hours type="weekly">1h40 por semana</hours>
                </studyHistory>

                <particActivities type="culture_PT">
                    <activity value="no">Não</activity>
                </particActivities>

                <holidayPeriod where="Portugal">
                    <freq value="2" type="yearly">Duas vezes por ano</freq>
                    <langUsage lang="PT" value="always">Sempre</langUsage>
                </holidayPeriod>
            </teiHeader>
        </TEI>

        <TEI>
            <!–Metainformação da narrativa 2–>
        </TEI>
    </teiCorpus>
```

4.5. LIMITAÇÕES DA ANOTAÇÃO AUTOMÁTICA

Embora possa existir uma variedade de tipos e formas de anotação, como já fora referido, a anotação linguística é normalmente a mais comum, dado que permite inserir informação linguisticamente relevante nos corpora. Gries e Berez (2017) expõe os vários tipos de formatos de anotação, sendo o mais recorrente aquele que se designa por anotação em linha ou incorporada (do inglês '*inline or embedded annotation*') (p. 393). Neste formato a anotação de um corpus ocorre no mesmo ficheiro e na mesma linha dos dados originais do corpus que está a ser anotado (Gries & Berez, 2017, p. 393). Segundo as autoras, este formato é bastante utilizado para a

lematização e etiquetação morfossintática do corpus. Gries e Berez (2017) referem, ainda, a anotação multinível ou interlinear (do inglês '*multi-tiered or interlinear annotation*'), onde os dados primários do corpus e a sua anotação estão no mesmo ficheiro, mas em linhas diferentes e, ainda, a existência de formatos em que os dados primários do corpus e a sua anotação são armazenados em ficheiros ou estruturas de dados separados (p. 393).

Relativamente à forma em que os corpora são anotados, o processo é quase sempre automatizado e com recurso a *softwares* treinados para o efeito, pois os corpora contemporâneos exigem o manuseamento de ferramentas para o seu processamento devido à sua grande dimensão (Newman & Cox, 2020, p. 26). Porém, o processo de anotação dos corpora pode ser semiautomatizado e, em casos mais específicos, pode ser feito manualmente (Granger et al., 2002, p. 16). Segundo Granger *et al*. (2002), a etiquetação morfossintática do corpus é um bom exemplo de anotação totalmente automática, uma vez que o etiquetador atribui a cada palavra do corpus uma etiqueta que indica a sua pertença a uma classe de palavras (p. 16). Apesar desta ser a mais comum atualmente, importa referir que a tarefa de automatizar a anotação de níveis mais elevados de processamento linguístico (por exemplo, categorias semânticas, pragmáticas ou discursivas) é bastante complexa e varia dependendo da língua em causa (Hovy & Lavid, 2010, p. 14). Por seu turno, um editor de erros é um exemplo de uma ferramenta de anotação semiautomática, que permite aos investigadores assinalar erros num texto (Granger et al., 2002, p. 16). Granger *et al*. (2002) esclarecem, também, que qualquer característica linguística pode ser anotada através de etiquetas criadas para um determinado objetivo de investigação e introduzidas manualmente no corpus (p. 16).

Não obstante, e como referido por Hovy e Lavid (2010), a anotação, seja ela automática ou manual, tem as suas limitações. Por conseguinte, é preciso ter em consideração que a maioria das ferramentas utilizadas para o processamento dos corpora são treinadas através de um conjunto de dados linguísticos em grande quantidade e de uma variedade de géneros e estilos. Normalmente, esses dados provêm de fontes cuja língua adquire um carácter mais cuidado (como nos livros, artigos e/ou jornais), de forma a garantir consistência e qualidade no treino. Segundo alertam Newman e Cox (2020) quanto mais o corpus diferir lexical e estruturalmente do tipo de textos que serviu de base para o treino das ferramentas de anotação, mais expectável será que a precisão da anotação diminua (p. 36). Isto acontece porque as ferramentas de anotação são sensíveis ao contexto e às características linguísticas para as quais foram treinadas. Desta forma,

textos com diferentes estruturas, géneros, estilos ou até mesmo variedades linguísticas podem apresentar padrões únicos que não foram adequadamente capturados durante o processo de treino das ferramentas. Assim, quanto maior a discrepância entre o texto a ser anotado e o(s) corpus/corpora original de treino, maior a probabilidade de ocorrerem erros de anotação. Por exemplo, os corpora de aprendizagem podem apresentar dificuldades no que diz respeito à anotação automática, dado que os informantes estão, ainda, a adquirir o léxico e as estruturas de uma determinada língua (Newman & Cox, 2020, p. 36). Também no seu artigo, Panunzi *et al.* (2004) destacou problemas enfrentados na anotação automática do discurso espontâneo em italiano, que incluía gírias, regionalismos, onomatopeias, entre outros elementos linguísticos, que provocaram falhas no processo de anotação. É, por isso, recomendada uma escolha cuidadosa das ferramentas para o processamento de corpora e uma revisão minuciosa dos resultados obtidos.

4.6. NORMALIZAÇÃO ORTOGRÁFICA

De acordo com Granger *et al.* (2002), um corpus de aprendizagem anotado deve basear-se idealmente num *software* de anotação normalizado, a fim de assegurar a comparabilidade de corpora de aprendizagem com os corpora nativos anotados (p. 10). No entanto, a normalização ortográfica (em inglês '*spell checker*') de corpora é sempre um desafio. Se, por um lado, é necessário normalizar para efeitos de processamento eletrónico, por outro, pretende-se manter o texto no seu formato original produzido pelo aprendente. Del Río *et al.* (2016) acrescentam que este tipo de dados produzidos pelos informantes são "extremely useful to assess, for instance, the difficult areas for the learning process according to the student's L1, the discourse restructuring or errors triggered by homophone words" (p. 10). Desta forma, Granger *et al.* (2002) alertam para a complexidade dos dados recolhidos a este grupo específico de informantes, pois os erros ortográficos são muito comuns, e por isso, exige o desenvolvimento de ferramentas como *softwares* de marcação de erros (em inglês '*error tagging software*') (p. 10). Atente-se no seguinte exemplo retirado do corpus em estudo (figura 7):

Figura 7

Excerto de uma narrativa retirado do subcorpus em PE.

A Elefantina também queria brincar com o avião
e então disidio tiralo ao Girafo.
O Girafo chateouse,
mas a Elefantina continua a brincar
até que o descheu cair na água
o Girafo ficou tão sangado
que comeseu a gritar com a Elefantina tão alto
que ela ficou com medo.

Para efeitos de marcação de erros, pode também ser fornecida uma versão normalizada (ortográfica, lexical ou sintática) para cada *token* (Del Río et al., 2016, p. 10). De acordo com Río *et al.* (2016), uma vez que os erros dos aprendentes afetam a etiquetação automática do POS e a lematização, o POS e o lema por defeito são normalizados, ou seja, corrigidos quando necessário e armazenados no primeiro nível de anotação de erros (ortográfico) (p. 10). No caso de corpora de aprendizagem, é importante manter os dados na sua forma original, mas acrescentar uma camada de informação que tem como função normalizar o erro ortográfico. Por sua vez, este processo vai contribuir para otimizar os processos de lematização e de etiquetação morfossintática e assim, através deste procedimento de controlo manual, garantir que as análises subsequentes se baseiem em informações corrigidas e mais precisas, tomando em consideração os erros cometidos pelos informantes.

No seu artigo, Del Río *et al.* (2016) exploram questões relativas à produção ortográfica, através da criação de um corpus de aprendizagem escrito e oral – *COPLE2*[52] – produzido por crianças de diferentes L1, tendo o PE como sua L2. Para tal, o corpus foi processado eletronicamente, aos níveis morfossintático, ortográfico e lematizado. Através dos resultados, as autoras classificam a etiquetação de erros como uma "mais-valia" nos corpora de aprendizagem, dado que este recurso, depois de anotado, permitiu a extração de dados quantitativos (estatísticas de erros) e qualitativos (tipos de erros), que permitiram, ainda, evidenciar as dificuldades dos aprendentes. Como referido, esta é uma abordagem bastante comum entre investigadores quando trabalham com corpora de aprendizagem. No âmbito do projeto desta dissertação não se procede à etiquetação de erros, mas apenas ao processamento eletrónico morfossintático e à lematização

[52] http://teitok.clul.ul.pt/cople2/

do corpus, com o objetivo de verificar o comportamento das ferramentas em relação a este tipo de corpora e desta forma extrair dados sobre a produção escrita dos informantes. Não obstante, sugere-se, para futuros trabalhos de investigação, um controlo manual dos resultados produzidos pelas ferramentas de processamento e, também, a etiquetação de erros.

4.7. LEMATIZAÇÃO

De acordo com Gries e Berez (2017), a lematização (do inglês '*lemmatization*') constitui um dos tipos mais básicos de anotação, cujo processo corresponde à identificação e marcação de cada palavra num corpus com a sua forma de base – o lema (p. 383). Durante vários séculos, o termo lema (do inglês '*lemma*') era utilizado em disciplinas como a lógica, filosofia e matemática (Dash, 2021, p. 166). Não obstante, este termo mudou de sentido na linguística e na lexicografia, onde viria a adquirir um novo sentido para se referir a "an entry word or a headword of a dictionary" (Dash, 2021, p. 167).Também Newman e Cox (2020) acrescentam "in lemmatization, each orthographic word encountered in a corpus is assigned a lemma, or 'base form', which provides a level of abstraction from any inflection that might appear in the original orthographic word." (p. 30). Assim, através da lematização do corpus, é possível simplificar o processo de extração e análise de informação, agrupando diferentes formas da mesma palavra sob uma única representação sem ter em conta as suas propriedades gramaticais (por exemplo, tempo verbal, modalidade, número ou género) (Biber et al., 1998, p. 94). Na tabela 8, é possível observar que palavras como *gatos*, *gata*, *gatas*, *gatinhos* ou *gatinhas* serão agrupadas sob uma forma base única, que neste caso seria "gato".

Tabela 8

Exemplificação do processo de lematização dando origem ao lema "gato".

Palavras flexionadas	*gatos*, *gata*, *gatas*, *gatinhos*, *gatinhas*
lema	flexões
"gato"	*-s* *-a* *-as* *-inhos* *-inhas*

Segundo Proisl *et al.* (2020), a lematização é crucial para fins gerais de indexação do corpus, bem como para muitas aplicações na lexicografia, classificação de textos ou na análise de discurso (p. 6142). Por conseguinte, os lemas correspondem à forma canónica da palavra (em inglês '*headwords*') encontrados nos dicionários. Newman e Cox (2020) explicam que "like the lemmas found in corpora, dictionary headwords often aim to represent a base word form, rather than provide separate entries for each distinct inflected word form" (p. 30). Tal como a lematização permite agrupar várias formas da mesma palavra base, as etiquetas semânticas agrupam vários sentidos de palavras relacionadas, que também podem ser exploradas, por exemplo, para a análise do discurso (Proisl et al., 2020, p. 6142). O processo de lematização tem como finalidade produzir uma lista separada de lemas e as suas possíveis flexões (Dash, 2021, p. 166). Desta forma, o lema serve para agrupar formas flexionadas, ortográficas ou fónicas, juntando palavras estruturalmente transformadas, mas gramaticalmente relacionadas sob uma forma base (Granja & Romero, 2015, p. 855). Assim, tanto os vocábulos nos dicionários como os lemas encontrados nos corpora têm um objetivo semelhante, que é o de permitirem aos investigadores localizar mais facilmente a informação. Newman e Cox (2020) afirmam que "searches based on lemmas can be invaluable when working with corpora of languages with rich inflection, such as Romance languages like French or Spanish, where a single regular verb may have dozens of distinct inflected forms" (p. 30).

Por outro lado, "o lematizador baseia-se num léxico computacional disponível para cada língua" (Gamallo & Garcia, 2017, p. 21), requerendo uma base de dados lexical representativa, que seja *standard*, *Unicode compatible*, sem erros, legível por máquina, normalizada, processada e formatada (Dash, 2021, p. 166). Estas características são indispensáveis para o desempenho da lematização, visto que o manuseamento de textos não formatados e não normalizados são suscetíveis à produção de falhas e obstáculos no seu processamento. Por esse motivo, e como mencionado no ponto 4.7., a normalização do texto é de extrema importância, a fim de corrigir e evitar possíveis incongruências ortográficas e textuais presentes no corpus. Segundo Dash (2021), o processo de lematização começa depois da normalização ortográfica e opera apenas em palavras flexionadas para produzir listas separadas de lemas e inflexões (p. 166). A base de dados de lemas é utilizada na análise morfológica, aprendizagem automática, ensino de línguas, compilação de dicionários e outros campos de linguística baseada em aplicações (Dash, 2021, p. 166).

4.8. ETIQUETAÇÃO MORFOSSINTÁTICA

A etiquetação morfossintática (em inglês '*Part-of-speech tagging*' ou '*POS tagging*') é um dos tipos de anotação mais frequentes e mais explorados quando se trabalha com corpora, visto que apresenta resultados relevantes para muitos estudos linguísticos de corpus e porque facilita muitos outros processos de anotação, como a lematização, a análise sintática, a anotação semântica, entre outros (Gries & Berez, 2017, p. 383). A etiquetação de um corpus constitui uma parte primordial nas pesquisas com corpora, uma vez que possibilita que o corpus cumpra o seu papel de instrumento para a investigação (Othero & Ayres, 2014, p. 46). De acordo com Armstrong *et al*. (1999), "Part-of-speech tagging consists of assigning to a word its disambiguated part of speech tag in the sentential context in which this word is used" (p. 8). Desta forma, é atribuída a cada palavra do corpus uma anotação correspondente à sua classe gramatical, indicando a sua função gramatical naquela frase específica. As etiquetas codificam geralmente categorias sintáticas como substantivos, verbos, adjetivos, entre outras, bem como subcategorias selecionadas e características morfológicas como número e tempo verbal (Brants, 2006, p. 221). De acordo com Brants (2006), os conjuntos de etiquetas *standard* contêm 20 a 100 categorias e, em alguns casos, abrangem 400 ou mais categorias (p. 221). Por conseguinte, a tarefa de um etiquetador morfossintático é distinguir entre os diferentes usos que as palavras podem adquirir (Brants, 2006, p. 221) mediante o contexto em que são utilizadas, pois estas podem adquirir várias funções gramaticais. Atente-se nas seguintes frases[53]:

a) A ***luta*** pelos direitos das mulheres jamais será esquecida.

b) O Pedro ***luta*** contra uma doença rara.

Em ambas as frases a palavra "luta" é utilizada, porém, e embora sejam semelhantes quanto à grafia e à fonética, pertencem a classes gramaticais distintas. Na frase (a), à palavra "luta" foi-lhe atribuída a etiqueta 'NCFS' (*Nome comum feminino singular*). Já na frase (b), a mesma palavra constitui o *verbo principal* da oração – 'VMI' – no modo indicativo. Atente-se nas tabelas 9 e 10 que mostram a etiquetação morfossintática completa de ambas as frases.

[53] As frases do exemplo foram etiquetadas com o etiquetador *TreeTagger* (Schmid, 1994), disponível em http://corpora.lancs.ac.uk/tree-tagger/.

Tabela 9

Etiquetação morfossintática do exemplo (a) com o etiquetador TreeTagger.

Token	POS-Tag	Lema	Designação
A	DA0	o	
luta	**NCFS**	**luta**	**Nome comum feminino singular**
pelos	SP+DA	por+os	
direitos	NCMP	direito	
de	SPS	de	
as	DA0	o	
mulheres	NCFP	mulher	
jamais	RN	jamais	
será	VMI	ser	
esquecida	VMP	esquecer	
.	Fp	.	

Tabela 10

Etiquetação morfossintática do exemplo (b) com o etiquetador TreeTagger .

Token	POS-Tag	Lema	Designação
O	DA0	o	
Pedro	NP0	pedro	
luta	**VMI**	**lutar**	**Verbo Principal Indicativo**
contra	SPS	contra	
uma	DI0	um	
doença	NCFS	doença	
rara	AQ0	raro	
.	Fp	.	

Nesse sentido, e como mencionado acima, é notório o papel desambiguador do etiquetador morfossintático durante o processo de etiquetação das palavras do corpus. Um dos principais desafios para os etiquetadores automáticos é a desambiguação. A qualidade dos resultados do processo de anotação morfossintática depende da qualidade do *input* utilizado para treinar o etiquetador, evidenciado a importância do fator língua. O seguinte exemplo mostra o primeiro parágrafo retirado de uma das narrativas escritas em AP.

Eines Tages gingen Giraffo und Elefantina in das nahe gelegene Schwimmbad. Giraffo hatte ein Spielzeug dabei, ein Flugzeug. Er fing an damit zu spielen. Elefantina schaute ihm beeindruckt zu. Doch irgendwann wurde sie neidisch auf ihren Freund und nahm ihm das Flugzeug einfach aus der Hand.

Este excerto foi etiquetado morfossintaticamente utilizando a ferramenta *TagAnt*[54] (Anthony, 2015), um *software* bastante intuitivo destinado à lematização e à etiquetação morfossintática através do pacote de anotação (*annotation package*) *TreeTagger*[55] (Schmid, 1994). Atente-se no mesmo excerto etiquetado, abaixo representado:

> Eines_**ART** Tages_**NN** gingen_**VVFIN** Giraffo_**NN** und_**KON** Elefantina_**NE** in_**APPR** das_**ART** nahe_**ADJA** gelegene_**ADJA** Schwimmbad_**NN** ._**$.** Giraffo_**NE** hatte_**VAFIN** ein_**ART** Spielzeug_**NN** dabei_**PROAV** ,_**$,** ein_**ART** Flugzeug_**NN** ._**$.** Er_**PPER** fing_**VVFIN** an_**APPR** damit_**PROAV** zu_**PTKZU** spielen_**VVINF** ._**$.** Elefantina_**NE** schaute_**VVFIN** ihm_**PPER** beeindruckt_**VVPP** zu_**PTKVZ** ._**$.** Doch_**KON** irgendwann_**ADV** wurde_**VAFIN** sie_**PPER** neidisch_**ADJD** auf_**APPR** ihren_**PPOSAT** Freund_**NN** und_**KON** nahm_**VVFIN** ihm_**PPER** das_**ART** Flugzeug_**NN** einfach_**ADV** aus_**APPR** der_**ART** Hand_**NN** ._**$.**

Através do *TagAnt*, é, também, possível a visualização dos resultados obtidos de forma vertical, tornando a sua leitura mais intuitiva (tabela 11). Na primeira coluna da tabela 11, observa-se cada *token* do parágrafo, seguida da etiqueta morfossintática resultante do processamento do excerto. Foi, ainda, adicionada uma terceira coluna com a forma base (lema) correspondente a cada *token*. Para consulta das etiquetas, recomenda-se o acesso à ligação disponível em rodapé[56].

Tabela 11

Processamento do excerto retirado do subcorpus em AP através da ferramenta TagAnt.

Token	POS-Tag	Lema
Eines	ART	ein
Tages	NN	Tag
gingen	VVFIN	gehen
Giraffo	NN	Giraffo
und	KON	und
Elefantina	NE	Elefantina
in	APPR	in
das	ART	der
nahe	ADJA	nah
gelegene	ADJA	gelegen
Schwimmbad	NN	Schwimmbad
.	$.	–
Giraffo	NE	Giraffo
hatte	VAFIN	haben

[54] https://www.laurenceanthony.net/software/tagant/
[55] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[56] https://www.sketchengine.eu/german-stts-part-of-speech-tagset/

ein	ART	ein
Spielzeug	NN	Spielzeug
dabei	PROAV	dabei
,	$,	–
ein	ART	ein
Flugzeug	NN	Flugzeug
.	$.	–
Er	PPER	er
fing	VVFIN	fing
an	APPR	an
damit	PROAV	damit
zu	PTKZU	zu
spielen	VVINF	spielen
.	$.	–
Elefantina	NE	Elefantina
schaute	VVFIN	schauen
ihm	PPER	ihm
beeindruckt	VVPP	beeindrucken
zu	PTKVZ	zu
.	$.	–
Doch	KON	doch
irgendwann	ADV	irgendwann
wurde	VAFIN	werden
sie	PPER	sie
neidisch	ADJD	neidisch
auf	APPR	auf
ihren	PPOSAT	ihr
Freund	NN	Freund
und	KON	und
nahm	VVFIN	nehmen
ihm	PPER	ihm
das	ART	der
Flugzeug	NN	Flugzeug
einfach	ADV	einfach
aus	APPR	aus
der	ART	der
Hand	NN	Hand
.	$.	–

Neste excerto, verifica-se que o resultado do processo de etiquetação automático foi relativamente bem sucedido, uma vez que a produção escrita por parte do informante apresenta um nível avançado de competência escrita no AP. No caso de narrativas que apresentem mais erros gramaticais, o resultado do processo de etiquetação automático não apresentará este nível de sucesso.

4.8.1. Conjuntos de etiquetas morfossintáticas

Segundo Schmid (1999), grande parte do trabalho desenvolvido sobre etiquetação morfossintática concentrou-se na língua inglesa, uma vez que já existia muito material de treino com etiquetas manuais para o inglês e os resultados podiam ser comparados com os de outros investigadores (p. 13). Desta forma, partiu-se do princípio de que os métodos desenvolvidos para o inglês também

funcionariam noutras línguas (Schmid, 1999, p. 13). Porém, essa premissa viria a cair por terra durante o desenvolvimento de etiquetadores para outras línguas. De acordo com o autor, um dos problemas decorre da produtividade morfológica das línguas, que, no caso da língua alemã, resulta num grande número de formas diferentes de palavras, o que, por sua vez, conduz a um grande número de parâmetros lexicais (Schmid, 1999, p. 13). Outra problemática detetada pelo autor refere-se à falta de corpora de grandes dimensões que estejam anotados de forma fiável para efeitos de treino (Schmid, 1999, p. 13). Desta forma, a grande variedade de possibilidades entre línguas, dá origem à criação de vários conjuntos de etiquetas, onde cada um refletirá as características morfossintáticas da sua própria língua (Armstrong et al., 1999, p. 8).

As etiquetas são escolhidas a partir de um conjunto pré-determinado, tais como categorias morfossintáticas (substantivo, verbo, preposição, etc.) e incluindo, entre outras, a declinação (por exemplo, nominativo, genitivo), a pessoa, o número ou o género (por exemplo, terceira pessoa, singular, feminino), o modo e o tempo (por exemplo, subjuntivo, imperfeito) (Armstrong et al., 1999, p. 8).

Na língua alemã, o *Stuttgart-Tübingen-Tagset* (Schiller et al., 1999), frequentemente abreviado como STTS, evoluiu ao longo dos anos até se tornar no conjunto de etiquetas padrão para a anotação morfossintática em alemão (Neunerdt et al., 2013, p. 142). Foi desenvolvido como parte do *TC Project*[57] pelas universidades de Stuttgart e Tübingen, mais especificamente pelo *Institut für maschinelle Sprachverarbeitung der Universität Stuttgart* e no *Seminar für Sprachwissenschaft der Universität Tübingen*, com o objetivo de padronizar e sistematizar a etiquetação gramatical das palavras em alemão (Neunerdt et al., 2013, p. 142). O STTS consiste num conjunto de 54 etiquetas morfossintáticas estruturado de forma hierárquica (Westpfahl, 2014, p. 2), que permite a análise computacional e a criação de corpora anotados para pesquisa linguística e PLN. Por sua vez, o conjunto de etiquetas é composto por 11 classes principais (refira-se a título exemplificativo ADJ para adjetivos), que se subdividem em subtipos (por exemplo, ADJD para adjetivo predicativo) (Zinsmeister et al., 2014, p. 4100). Observe-se a figura 8.

[57] https://www.ims.uni-stuttgart.de/en/research/projects/textkorpora-werkzeuge/

Figura 8

Representação das classes principais do conjunto de etiquetas STTS. Adaptado de Schiller et al. (1999).

1. Nomina (N)
2. Verben (V)
3. Artikel (ART)
4. Adjektive (ADJ)
5. Pronomina (P)
6. Kardinalzahlen (CARD)
7. Adverbien (ADV)
8. Konjunktionen (KO)
9. Adpositionen (AP)
10. Interjektionen (ITJ)
11. Partikeln (PTK)

Como mencionado por Zinsmeister *et al.* (2014), através dos subtipos é possível capturar nuances gramaticais no corpus. Atente-se no excerto retirado do subcorpus em AP, que foi anotado ao nível morfossintático (tabela 12).

Tabela 12

Excerto retirada do subcorpus em AP etiquetado ao nível morfossintático através da ferramenta TagAnt.

Token	POS-Tag
Giraffo	NE
war	VAFIN
überglücklich	ADJD
,	$,
weil	KOUS
er	PPER
sein	PPOSAT
Spielzeug	NN
wieder	ADV
hatte	VAFIN
und	KON
Elefantina	NE
war	VAFIN
glücklich	ADJD
weil	KOUS
ihr	PPOSAT
Freund	NN
wieder	ADV
glücklich	ADJD
war	VAFIN
.	$.

Através dos resultados obtidos, é possível observar as várias categorias gramaticais resultantes do processo de etiquetação morfossintático. Um exemplo de uma das classes principais pode ser visível através da etiqueta "ADV" para Advérbio, atribuída ao *token* "wieder" ou através da etiqueta "NN" para Nome, atribuída ao *token* "Freund". Por outro lado, um exemplo de uma subcategoria pode ser observada através da etiqueta "ADJD" para adjetivo predicativo atribuído ao *token* "überglücklich".

Ainda sobre o processo de anotação, Zinsmeister *et al.* (2014) afirmam que "further morphosyntactic attributes can be added as a second layer that specializes the annotation on the basis of lexical information" (p. 4100). Por outras palavras, cada etiqueta pode ser representada por um par de "atributo-valor", o que ajuda a descrever as características da palavra de maneira mais detalhada, como por exemplo a atribuição dos género e número nos substantivos, ou do modo e tempo nos verbos. Esta estrutura organizada permite fazer distinções precisas entre diferentes tipos de palavras e suas funções na língua alemã. De acordo com Schiller *et al.* (1999), as etiquetas "orientieren sich am "TEI Starter Set Of Grammatical-Annotation Tags" mit Ausnahme der Kardinalzahlen, die durch den Wert *cardinal* beim Merkmal *numeral* der Adjektive abgedeckt werden und der Konjunktionen, die dort von den zwei Kategorien *subordinators* und *coordinators* reprasentiert werden" (p. 4). Este conjunto de etiquetas morfossintáticas é amplamente utilizado em projetos de análise linguística, como a análise sintática, a identificação de padrões morfossintáticos e a extração de informações linguísticas de textos na língua alemã.

CAPÍTULO V – PROCESSAMENTO DO CORPUS EM ESTUDO

5.1. DESAFIOS DOS CORPORA DE APRENDIZAGEM PARA O PROCESSAMENTO ELETRÓNICO

Realizar tarefas de PLN em corpora de aprendizagem pode ser um desafio, visto que os textos são produzidos por aprendentes de uma língua, e como tal, constituem um grupo de falantes com características particulares. Segundo van Rooy (2015), um corpus de aprendizagem oferece algumas vantagens quando comparado com as fontes de dados tradicionais utilizadas na investigação sobre a aquisição de uma segunda língua e o ensino de línguas estrangeiras, nomeadamente quanto à sua dimensão, variabilidade (podem ser incluídos mais indivíduos e uma gama mais vasta de tipos de texto) e automatização de vários aspetos da análise de dados (p. 79). Em contrapartida, existem, também, vários projetos de investigação que apontam para os desafios associados à anotação de corpora de aprendizagem (van Rooy, 2015, p. 103). Por conseguinte, o autor sugere que "Learner corpus researchers should therefore use annotation with awareness of the limitations, including the assumptions that are made when annotating data with tools that have been designed for native-speaker data in the first instance" (van Rooy, 2015, p. 103).

Até à data, os tópicos mais abordados na investigação de corpora de aprendizagem são a identificação e anotação dos tipos de erros cometidos pelos informantes que produzem os corpora (Cardoso et al., 2014; Gries & Berez, 2017; Mendes, 2016). A presença de erros nos corpora de aprendizagem pode derivar de vários fatores, podendo estes estarem relacionados com os diferentes contextos linguísticos, níveis de proficiência ou pelos diferentes contextos de aprendizagem dos informantes (Mukherjee & Götz, 2015, p. 423). Essa instabilidade torna difícil o desenvolvimento de esquemas ou modelos de anotação de erros exatos e abrangentes, sendo considerado por van Rooy (2015) a principal e mais óbvia razão pelo fraco desempenho dos sistemas de anotação automática em corpora de aprendizagem (p. 89). Os lematizadores são ferramentas que utilizam o método de processamento da língua natural e, por conseguinte, a compreensão do computador depende em grande medida da qualidade da configuração da morfologia, sintaxe, semântica, fonética e gramática inserida no sistema (Kharis et al., 2021, p. 190). Por conseguinte, e a fim de assegurar consistência da anotação do corpus, é necessária uma retificação redobrada devido à falta de concordância entre anotadores, que podem diferir nas suas interpretações dos erros, resultando em incongruências no processo de anotação (Lüdeling & Hirschmann, 2015, p. 149). Os corpora de aprendizagem também apresentam frequentemente

ambiguidade linguística (Leacock et al., 2015, p. 573), o que torna difícil determinar, por exemplo, se uma expressão ou construção específica é um erro ou uma variação válida da própria língua. A ambiguidade é, como tal, uma das principais razões pelas quais o processamento da linguagem é difícil, uma vez que a língua é claramente ambígua em múltiplas formas.

Em segundo lugar, os corpora de aprendizagem destacam-se pela presença de elementos únicos, características das tentativas dos aprendentes de transferir padrões e estruturas da sua L1 para a sua L2 (Durrant & Siyanova-Chanturia, 2015, p. 62). Estas características interlinguísticas, que são influenciadas pela sua língua materna e por experiências anteriores de aprendizagem de línguas, podem comprometer a análise linguística e dificultar a aplicação de técnicas de processamento linguístico normalizadas. Com efeito, van Rooy (2015) aponta a segunda razão pela qual os anotadores automáticos desempenham ineficientemente a sua tarefa, pois estes foram desenvolvidos com base num tipo específico de dados, geralmente dados de falantes nativos provenientes de registos como a escrita de jornais (p. 89).

Também a presença de elementos ou estruturas gramaticais em algumas línguas podem resultar em desafios para os anotadores. No alemão, bem como em todas as línguas germânicas (Dehé, 2015, p. 611), os verbos de partícula separável (em alemão '*Trennbare Verben*') constituem um obstáculo no domínio do PLN, sendo inclusive classificados por Sag *et al.* (2002) como "a pain in the neck" (p. 1). Foi demonstrado que este tipo de verbo representa um sério problema para a tecnologia da linguagem, devido à sua ambiguidade e ao seu comportamento aparentemente imprevisível em termos de significado (Sag et al., 2002, p. 6). Desta forma, as partículas podem facilmente ser confundidas com preposições ou vice-versa. Tenha-se por exemplo o verbo *anfangen* (em português 'começar'), constituído pela forma base "fangen" e a partícula separável "an".

DE *Heute* ***fange*** *ich* ***an*** *zu arbeiten.*

PT *Hoje começo a trabalhar.* [Tradução livre]

Ao contrário das línguas românicas, a forma base do verbo *anfangen* foi conjugada na frase de acordo com a primeira pessoa do singular (*fange*) e a partícula "an" colocada na posição correspondente. Porém, as partículas resultantes dos *Trennbare Verben* podem ocorrer numa

posição muito mais afastada da forma base ou até mesmo no final da frase, e dessa forma, criar ambiguidade ao *software* no momento do processamento. Atente-se no seguinte exemplo:

DE *Heute* ***kommt*** *João von seiner Reise nach Madrid mit seinen Freunden* ***zurück***.

PT *O João regressa hoje da sua viagem a Madrid com os seus amigos*. [Tradução livre]

O verbo "zurückkommen" (em português 'regressar') é um verbo de partícula separável composto pela forma base "kommen" e a partícula separável "zurück". Como se pode observar neste caso, a partícula ocorre no final da frase, estando suscetível de ser confundida com um advérbio, por exemplo. Por esse motivo, os *Trennbare Verben* requerem um tratamento especial no domínio do PLN.

Pese embora esta amálgama de obstáculos às ferramentas de processamento linguístico, a verdade é que a presença deste tipo de elementos, como os erros, conferem riqueza aos corpora, fornecendo informações valiosas sobre o processo de aquisição das línguas, bem como uma visão global do processo de aprendizagem das mesmas. Ademais, é com base nestes elementos que muitas ferramentas de processamento do texto podem ser treinadas e melhoradas.

5.2. FERRAMENTA DE ANOTAÇÃO E ANÁLISE DO CORPUS: SKETCH ENGINE

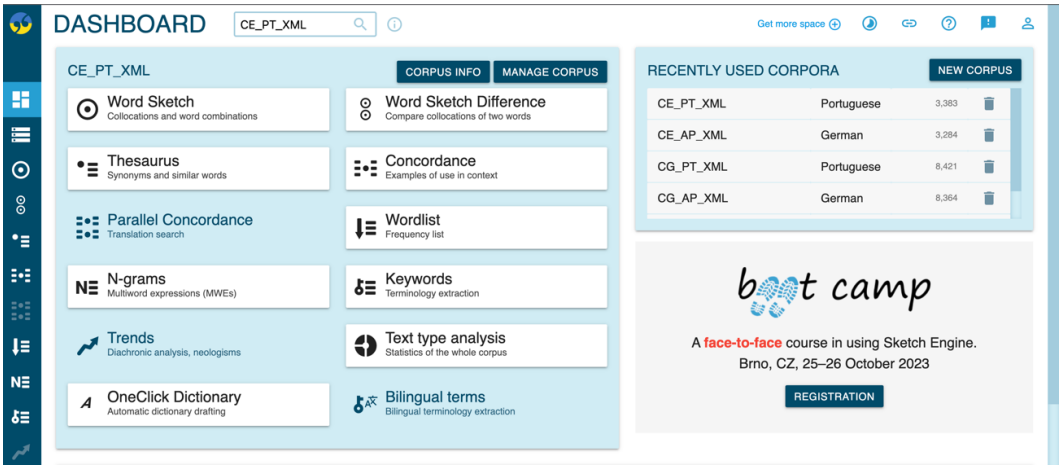
O método de anotação totalmente automática (*automatic tagging*) é o mais eficiente em termos de tempo e de custos, onde o *software* atribui etiquetas aos dados automaticamente, sem interagir com o utilizador (van Rooy, 2015, p. 86). Este método é frequentemente utilizado na etiquetação morfossintática. Rooy (2015) sugere, ainda, uma opção de etiquetação intermédia, denominada de etiquetação interativa (*interactive tagging*), na qual parte do trabalho é feito por um programa de computador, mas o utilizador tem de tomar certas decisões ou inserir certos marcadores e uma terceira opção de anotação, que seria totalmente manual (*manual annotation*) (p. 87). Esta última é desempenhada pelo utilizador, onde este insere todas as etiquetas diretamente no corpus.

O processamento deste corpus foi desempenhado pelo sistema de gestão de corpora *Sketch Engine* (Kilgarriff et al., 2014), um *web-based program* que disponibiliza uma série de

funcionalidades para a criação, anotação e análise linguística do corpus. Esta ferramenta de gestão de corpora, amplamente utilizada na lexicografia, ensino de línguas, tradução, entre outras disciplinas, foi concebida pelo linguista, lexicógrafo e investigador Adam Kilgarriff em 2003. Inicialmente, o *Sketch Engine* foi desenvolvido para ser manuseado por lexicógrafos na compilação de dicionários, contudo, esta ferramenta tem vindo a ser amplamente utilizada em PLN devido às suas sofisticadas e variadas funcionalidades (Bonial et al., 2013, p. 2). De acordo com Albi (2019), "Sketch Engine is one of the most versatile, complete and user-friendly corpus tools that exists today" (p. 33). A autora acrescenta que "it contains some of the biggest monolingual and parallel corpora in many languages and allows users to easily create their own ad-hoc corpus. Furthermore, it offers a comprehensive set of search and concordance features in an easy-to-use interface" (Albi, 2019, p. 33). Hu e Yang (2015) acrescentam que o *Sketch Engine* "refers to two different things: the software, and the web service. The web service includes, as well as the core software, a large number of corpora pre-loaded and 'ready for use', and tools for creating, installing and managing users' own corpora" (p. 32). O *Sketch Engine* disponibiliza corpora em diferentes línguas, desde o árabe ao inglês, grego, russo, vietnamita, galês, entre muitas outras, constituindo uma ferramenta muito interessante e dinâmica para explorar o funcionamento da língua. Esta dispõe de uma série de funcionalidades, tais como: *Thesaurus*, *Wordlist*, *Concordance*, *N-grams*, *Word Sketch*, *Sketch-Difference*, entre outras. Através da ilustração 4, pode ter-se uma perspetiva geral do aspeto e das funcionalidades oferecidas por esta ferramenta.

Ilustração 4

Painel das funcionalidades disponibilizadas pelo Sketch Engine[58].



[58] Acedido em outubro de 2023.

Não obstante, importa referir que o *Sketch Engine* é uma plataforma paga, que oferece apenas um período experimental de 30 dias. Contudo, atualmente o acesso pode ser gratuito e ilimitado caso o utilizador esteja vinculado a uma instituição de ensino superior que tenha subscrito o acesso ao *Sketch Engine* para os seus alunos. Para além disso, o facto de ser uma plataforma que funcione exclusivamente *online*, pode constituir alguns constrangimentos aos utilizadores, uma vez que requer o acesso a internet para poder aceder ao corpus e proceder à sua análise. Por outro lado, a sua natureza facilita o aceso ao corpus a partir de qualquer computador ou dispositivo eletrónico.

Das várias funcionalidades disponibilizadas para análise de corpora, as mais pertinentes neste projeto de dissertação são as *Wordlist* e a *Concordance*, pois de acordo com Hunston (2002b), representam os métodos e abordagens mais utilizados na linguística de corpus (p. 38).

5.2.1. Lista de frequências

Segundo Adolphs (2006), "with the first computerized corpora in the 1960s it became possible to generate frequency lists automatically and these have since been used as the basis for a number of different research purposes" (p. 5). A *frequency list*, ou lista de frequências, permite a criação de uma lista de palavras do corpus, mostrando o número de ocorrências no texto (por exemplo, tabela 12). Miller (2020) acrescenta que "a fundamental statistic in assessing the saliency of any linguistic feature is *frequency of occurrence*, or, simply, the number of times a feature of interest occurs in a dataset" (p. 77). As listas de frequências podem ocorrer por ordem alfabética ou por ordem de frequência, de forma ascendente ou descendente (Sinclair, 1991, p. 31). No *Sketch Engine*, uma lista de frequências pode ser gerada através da funcionalidade *Wordlist*. Esta, por sua vez, pode gerar uma lista de palavras com os *tokens* presentes no corpus, as classes de palavras, lemas, prefixos ou sufixos. Por defeito, a *Wordlist* apresenta todas as palavras da maior para a menor frequência no corpus. Sinclair (1991) explica que "this kind of list is helpful as a quick guide to the way words are distributed in a text, and for a number of more specific purposes" (p. 30). O autor continua, mencionando que:

> A frequency list of word-forms is never more than a set of hints or clues to the nature of a text. By examining the list, one can get an idea of what further information would be worth acquiring: or one can make guesses about the structure of the text, and so focus an investigation. (Sinclair, 1991, p. 31)

Para além de ajudar a identificar o vocabulário-alvo para estudo, as listas de frequência têm sido utilizadas para ajudar os educadores e investigadores a compreender melhor as exigências lexicais dos usos da língua-alvo ou para avaliar o desenvolvimento do vocabulário pelos informantes (Miller, 2020, pp. 77-78).

5.2.2. Concordância

Por seu turno, a *Concordance* oferece uma variedade de opções de pesquisa, constituindo uma forma de apresentação dos dados linguísticos para facilitar a análise de corpora (Adolphs, 2006, p. 5). De acordo com Baker (2006), através de linhas de concordância é possível obter "all the occurrences of a particular search term in a corpus, presented within the context that they occur in; usually a few words to the left and the right of the search term" (p. 71). A esta definição, Sinclair (1991) acrescenta que "a *concordance* is a collection of occurrences of a word-form, each in its own textual environment. In its simplest form, it is an index. Each word-form is indexed, and a reference is given to the place of each occurrence in a text" (p. 32). Por outras palavras, uma concordância apresenta linhas de concordância em que a palavra pesquisada aparece destacado ao centro com o contexto à sua esquerda e à sua direita (ver figura 10). Este resultado é designado por *Key Word In Context* (KWIC), em português por "palavra-chave em contexto" (Wulff & Baker, 2020, p. 162), sendo que "the quality of evidence about the language which can be provided by concordances is quite superior to any other method" (Sinclair, 1991, p. 42). Evison (2010) refere que as concordâncias tanto são úteis para "hypothesis testing and for hypothesis generation" (p. 129). Não obstante, através do uso de linhas de concordância, uma pesquisa pode ser concebida para mostrar não as palavras procuradas, mas um conceito que frequentemente coincide com estas (Hunston, 2002b, p. 38). Hunston (2002b) afirma que a pesquisa e análise das linhas de concordância pode ser agilizada se estas estiverem ordenadas alfabeticamente (p. 40). Desta forma, "the lines that are like each other in some way appear next to each other" (Hunston, 2002b, p. 40). Para fins de investigação, a organização por ordem alfabética pode ser útil, uma vez que permite identificar de forma mais eficiente, que tipos de itens lexicais antecedem e sucedem a KWIC mais frequentemente. Dependendo da ferramenta a ser utilizada, existem, ainda, *concordancers* que permitem ao investigador efetuar pesquisas mais detalhadas, nomeadamente a pesquisa através de uma frase ou classes de palavras específicas (Hunston, 2002b, p. 41).

Para o processamento do corpus em análise nesta dissertação, serão apresentados, nas secções 5.3., 5.4., 5.5. e 5.6., os resultados obtidos nos processos de lematização e de etiquetação morfossintática. Os resultados são fruto de um processamento automático do corpus de aprendizagem, que ao ter permanecido inalterado, contém vários erros produzidos pelos informantes. Por conseguinte, os resultados obtidos refletem a interpretação feita pela ferramenta *Sketch Engine* aquando do seu processamento. Assim, proceder-se-á apenas a uma exposição e descrição sucinta dos resultados obtidos, sendo estes, em casos excecionais, alvo de uma análise um pouco mais detalhada, sempre que se justifique ou se identifique algum fenómeno linguístico de relevância.

5.3. LEMATIZAÇÃO DO SUBCORPUS EM PORTUGUÊS EUROPEU

O processo de lematização do subcorpus em PE deu origem à seguinte lista de palavras (tabela 13), que contém apenas 50 entradas e mostra os lemas com mais frequência. Segundo a plataforma *Sketch Engine*, a lista de lemas obtidos contém um total de 409 lemas (consultar Anexo I) no subcorpus em PE.

Tabela 13

Lista de 50 entradas com os lemas do subcorpus em PE e a respetiva frequência.

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	o	449	14	brincar	50	27	depois	26	40	dia	20
2	e	138	15	muito	46	28	mas	26	41	se	20
3	avião	134	16	estar	42	29	não	26	42	cair	20
4	elefantino	122	17	amigo	39	30	ajudar	25	43	girafito	18
5	girafo	96	18	tirar	36	31	querer	23	44	por	18
6	ele	90	19	seu	35	32	certo	23	45	deixar	18
7	com	89	20	ter	34	33	triste	21	46	chegar	18
8	um	84	21	para	32	34	rede	21	47	conseguir	17
9	que	80	22	tentar	31	35	piscina	21	48	olhar	16
10	a	74	23	começar	28	36	em+a	21	49	de+o	16
11	ficar	70	24	de+a	28	37	outro	21	50	ver	16
12	elefante	64	25	feliz	27	38	brinquedo	20			
13	de	57	26	a+o	26	39	ir	20			

Através desta lista observam-se algumas *stopwords* nos primeiros lugares, como *o*, que ocupa a posição 1 e *e*, que ocupa a posição 2, e assim por diante. Estas, ao não terem sido removidas no processo de tokenização e serem comumente utilizadas na escrita, ocorrem com bastante frequência. Além disso, palavras como "avião", "elefantino" ou "girafo" também ocupam os primeiros lugares, denotando a temática do enredo das narrativas produzidas pelos informantes. Seguem-se, também, alguns lemas que remetem para verbos utilizados na descrição de ações como "ficar", "tentar", "começar", entre outros, e "brincar", na posição 14, que dá pistas sobre a trama da própria narrativa. As palavras com a mesma frequência estão dispostas por uma ordem arbitrária, como é o caso dos lemas "outro", "em+a" e "pisicna", que ocorrem 21 vezes.

Em contrapartida, existe um total de 208 palavras diferentes (e consequentemente 208 lemas diferentes) que ocorrem apenas uma vez. De acordo com Baayen (2001), "the words which occur once only in a text are known as hapax legomena, from Greek *hapax*, 'once', and *legomenon*, 'read'" (p. 8). Importa lembrar que, uma vez que as narrativas mantiveram a sua forma original, preservando os erros ortográficos produzidos pelos informantes, parte dos *hápax legomenon*, gerados através da lista de frequência do subcorpus em PE, são lemas que foram criados pela ferramenta com base em *tokens* não corrigidos. Apesar disso, os lemas que contêm erros ortográficos no contexto deste projeto são de extrema importância, como já fora referido. Através da tabela 14, é possível visualizar apenas 50 entradas dos 208 *hápax legomenon* presentes no subcorpus em PE (consultar Anexo I).

Tabela 14

Lista de frequência com os alguns exemplos de hápax legomenon presentes no subcorpus em PE.

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	achar	1	14	amigos	1	27	arrepender	1	40	brinca-va	1
2	a+os	1	15	animar	1	28	atenção	1	41	brincalhão	1
3	acompanhar	1	16	animação	1	29	aulto	1	42	brinquar	1
4	adulto	1	17	apanar	1	30	aviar	1	43	cadinho	1
5	acondeseu	1	18	aparesse	1	31	aí	1	44	caio	1
6	adurar	1	19	aperceber	1	32	aõ	1	45	capaz	1
7	aguã	1	20	aprecebeu-se	1	33	bambar	1	46	cara	1
8	ainda	1	21	apreximou-se	1	34	bdo	1	47	censegoiu	1
9	ajudasnos	1	22	aproximou-ce	1	35	bena	1	48	chati-adissímo	1
10	alcance	1	23	aquilo	1	36	bolso	1	49	cheio	1
11	algum	1	24	arancou	1	37	bom	1	50	chomou-o	1
12	alcançar	1	25	arrastar	1	38	borda	1			
13	ali	1	26	asustada	1	39	braço	1			

Dado que as ferramentas de processamento para lematização de corpora são treinadas com base em dicionário (Korenius et al., 2004, p. 627), a ferramenta, ao não reconhecer o *token* e ao não identificar outros *tokens* semelhantes ou derivados, gerou, automaticamente, lemas baseados em palavras com erros ortográficos. Na tabela 14 é possível observar alguns desses lemas, como "acondeseu" (posição 5), "adurar" (posição 6) ou "ajudasnos" (posição 9).

A lista completa de *hápax legomenon* mostra que muitos dos erros ortográfico presentes no subcorpus em PE advêm da incorreta e/ou falta de acentuação nas palavras, como em "tambem" (*também*), "àgua" (*água*) ou "pérto" (*perto*). Muitos destes erros, como "brincedo" (*brinquedo*), "comessou" (*começou*) ou "enveja" (*inveja*), surgem pela homofonia de algumas consoantes e/ou vogais e que provocam dúvida na parte escrita. Outros exemplos como "shorar" (*chorar*), "trubsou" (*tropeçou*) e "pichina" (*piscina*) denotam, também, esta incerteza provocada pela falta de concordância absoluta de som-grafema, levando os informantes a escrever muitas das palavras tal e qual como as pronunciam.

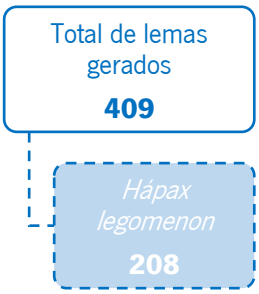
Foram, ainda, identificados erros ortográficos que evidenciam a transposição de regras gramaticais do AP para o PE, na medida em que muitos substantivos foram escritos com letra maiúscula, nomeadamente em "Elefante". Este fenómeno, bastante comum em falantes bilingues, designa-se, em linguística, de *cross-linguistic influence* ou transferência interlinguística a nível da ortografia. De grosso modo, a investigação em transferência interlinguística procura descrever o processo através do qual elementos de uma língua influenciam a aquisição ou a produção de outra língua em indivíduos bilingues (McManus, 2021, p. 1), ou seja, quando existe a transferência de propriedades linguísticas de uma língua para a outra durante a compreensão ou a produção. De acordo com Rajaa (2016), "numerous theories investigate its meaning, its occurrence, its effects on the process of SLA and the consequences of such a phenomenon for language learners" (p. 201). No domínio da aquisição de uma segunda língua, a investigação em transferência interlinguística é vasta, demonstrando, entre outras hipóteses, "how first language (L1) experience shapes second language (L2) learning" (McManus, 2021, p. 1). Desta forma, a aquisição da L2 é caracterizada por uma influência significativa da L1, especialmente nas fases iniciais de desenvolvimento (Montrul, 2015, p. 144). Contudo, também se observa o inverso, isto é a influência da L2 sobre a L1. Esta situação é verificada em especial em falantes de herança que são bilingues sucessivos, adquirindo a língua de herança como L1 e a língua maioritária do país de residência como L2. Sendo a L2 a língua principal de escolarização, o que se observa

frequentemente em falantes de herança é a influência da L2 sobre a L1/LH (Meir & Janssen, 2021, p. 3). Odlin (2005) acrescenta: "Crosslinguistic influence is an important topic not only for SLA research but also for studies of language contact, which usually emphasize the sociohistoric product of an acquisition process" (p. 4). A título exemplificativo, Wolters e Kim (2023) referem, no seu artigo, o par de línguas inglês-espanhol, que embora não derivem da mesma família, partilham algumas características segundo as autoras, como o alfabeto, a pronunciação semelhante de algumas consoantes, e ainda, a existência de cognatos, que são ortográfica e fonologicamente semelhantes (por exemplo, *class*, *clase*; *family*, *familia*; *photo*, *foto*; *rock*, *roca*) (p. 3). Desta forma, referem as autoras, "a dual language learner who starts to spell a word in a target language in which they are less proficient may employ the spelling pattern of the more acquired or salient language" (Wolters & Kim, 2023, p. 3). Dos casos observados nas narrativas, a transposição de certas regras gramaticais, como a escrita da primeira letra dos substantivos em maiúscula, foi bastante evidente, principalmente nos informantes de menor idade, onde a aquisição tanto do AP como do PE ainda se encontra em desenvolvimento. De acordo com Torrijos (2009), "the effects of cross-linguistic influences are not monolithic but instead vary considerably according to the social context of the language contact situation" (p. 148). Em contextos bilingues ou até mesmo multilingues, os efeitos da transferência interlinguística variam significativamente, sendo influenciados por vários fatores de índole social, como por exemplo o ambiente, as interações culturais ou as circunstâncias específicas em que ocorre o contacto linguístico e, também, por fatores individuais relacionados, por exemplo, com a idade do falante, a memória, estilos de aprendizagem, entre outros.

Relativamente ao processo de lematização do subcorpus em PE, contabilizou-se um total de 409 lemas e dos quais 208 correspondem a palavras que apenas ocorreram uma única vez (*hápax legomenon*). Atente-se no diagrama da figura 9.

Figura 9

Representação do total de lemas e hápax legomenon gerados no subcorpus em PE.



Por sua vez, a funcionalidade *Concordance* permite a visualização de um determinado item lexical tendo em conta o seu contexto no corpus, revelando, por exemplo, padrões de uso, significado e o comportamento deste no corpus. A figura 10 mostra a concordância do lema *ficar*, que foi o verbo com mais frequência no subcorpus em PE. Na figura é possível ver apenas algumas linhas de concordância (consultar Anexo II).

Figura 10

Concordância do lema "ficar" no subcorpus em PE.

Linhas	Left context	KWIC	Right context
1	stava zangado que começou a berrar com a Elefantina.</s><s>A Elefantina	ficou	com medo mas depois vieo um outro Elefant e a Elefantina dizlhe ajudasnos
2	anta com uma rede e tirou o avião da agua e deu o giraffo.</s><s>O giraffo	ficou	bem feliz e a Elefantina bambem porque estava a ver o seu amigo feliz.</s>
3	o avião ao girafo Depois a elefantina deixou o avião cair no piscina O girafo	ficou	tão bravo mesmo tão bravo que começou a grita com a elefantina.</s><s>O
4	segui tirar o avião la de dentro Depois ela devolveu o avião ao girafo No fim	ficaram	os dois felizes o girafo ficou feliz por ter o avião de volta e a elefantino por o
5	Depois ela devolveu o avião ao girafo No fim ficaram os dois felizes o girafo	ficou	feliz por ter o avião de volta e a elefantino por o girafo estar feliz Certo dia, a
6	ıpre a brincar com ele.</s><s>Corria de um lado para o outro e a Elefantina	ficava	a olhar.</s><s>Até que ela decidiu tirar o avião ao Girafo.</s><s>Ele ficou c
7	ficava a olhar.</s><s>Até que ela decidiu tirar o avião ao Girafo.</s><s>Ele	ficou	com medo, porque a Elefantina o podia deixar cair.</s><s>E foi isso que ac
8	afa muito brincalhona estava a brincar com o seu avião.</s><s>A Elefantina	ficou	com inveja.</s><s>E tirou lo avião.</s><s>quando ela estava a brincar com
9	ım o seu avião.</s><s>sem querer deixou o avião cair a pichina.</s><s>Ela	ficou	triste.</s><s>A girafa gritou com a Elefantina.</s><s>Depois apareceu outr
10	ı.</s><s>Depois veu a mãe com uma pá e tirou de la o avião.</s><s>Todos	ficarão	feliz.</s><s>A Mãe deu a girafa o avião.</s><s>A Elefantinaficou feliz de ve
11	lefantina trubsou e o aviá brincar com o avião caiu á água.</s><s>O Girafo	ficou	com tanta raiva que á Elefantina tava com medo á olhar para a cara do Gira

Através destes exemplos, é possível observar algumas das várias formas que o lema adquire, nomeadamente *ficou*, *ficaram*, *ficava* e *ficarão*, assim como os contextos em que foi utilizado. Torna-se, assim, evidente que a estrutura [**substantivo** + *ficar*], que ocorre 34 vezes, é a mais comum, visível, por exemplo, em "A **Elefantina** *ficou*". Atente-se na tabela 15. Foram, ainda, identificadas outras estruturas, tais como [*ficar* + **advérbio**] e [*ficar* + **substantivo**], ocupando as segunda e terceira estruturas com mais frequência no subcorpus em PE. Por oposição, a

estrutura que menos ocorre são construções perifrásticas do tipo [preposição + infinitivo], visível em "ficava a olhar" na 6 linha da figura 10.

Tabela 15

Estruturas com o lema "ficar" presentes no subcoprus em PE.

Estrutura	Exemplo	Frequência
substantivo + *ficar*	*o **girafo** ficou*	34
ficar + advérbio	*ficaram **muito** felizes*	29
ficar + substantivo	*ficaram **amigos***	19
sintagma preposicional (com/sem/por)	*ficou **com** inveja* *ficou **sem** ideias* *ficou triste **por** o seu avião*	16
e_ou	*ficou zangado **e** ralhou*	13
advérbio + *ficar*	***também** ficou*	4
preposição + infinitivo	*ficava **a olhar***	2

Wulff e Baker (2020) referem que uma das formas em que as linhas de concordância podem ser usadas de forma proveitosa é na identificação da prosódia semântica (p. 166). Isto é, através desta funcionalidade é possível perceber se um determinado lema, mediante o contexto em que se insere, está mais associado a aspetos positivos ou negativos, de acordo com o corpus em análise (Falchi, 2021, p. 20).

5.4. LEMATIZAÇÃO DO SUBCORPUS EM ALEMÃO PADRÃO

Através da lematização do subcorpus em AP, foram gerados um total de 436 lemas, 27 lemas a mais que no subcorpus em PE. Através da tabela 16, observam-se os 50 lemas mais frequentes no subcorpus em AP (consultar Anexo III).

Tabela 16

Lista de 50 entradas com os lemas do subcorpus em AP e a respetiva frequência.

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	sie	311	14	in	38	27	nicht	23	40	traurig	16
2	die	252	15	wollen	32	28	nehmen	23	41	auch	16
3	und	124	16	an	32	29	aber	23	42	hand	15
4	giraffo	114	17	wasser	30	30	da	23	43	so	15
5	elefantina	112	18	kommen	29	31	helfen	22	44	geben	15
6	flugzeug	111	19	dann	29	32	freund	21	45	fangen	15
7	eine	107	20	aus	27	33	dass	21	46	ihre	14
8	sein	103	21	spielzeug	27	34	tag	21	47	sagen	14
9	zu	88	22	seine	26	35	gehen	21	48	einfach	14
10	haben	69	23	wieder	25	36	wie	20	49	plötzlich	13
11	mit	59	24	sehr	25	37	weil	20	50	fallen	12
12	spielen	43	25	werden	25	38	was	19			
13	elefant	42	26	von	24	39	glücklich	19			

À semelhança do subcorpus em PE, são visíveis, através da lista da frequência (tabela 16), algumas *stopwords*, como "sie", "die" e "und" entre os lemas com mais frequência no texto. A primeira posição na tabela é ocupada pelo lema *sie*, que na língua alemã é utilizado tanto para se referir à terceira pessoa do singular (feminino) como à terceira pessoa do plural (masc./fem.). Este item lexical é, ainda, utilizado para a enunciação do registo formal nesta língua (por exemplo, *Haben Sie verstanden? / Percebeu?* [Tradução livre]). A forte utilização deste tipo de pronomes, evidenciam, também, o género literário em análise, pois como referido, o lema *sie* pode ser aplicado em diferentes contextos, e desta forma, utilizado para se referir a distintas pessoas e entidades na produção de diálogos e narrativas. As posições subsequentes são ocupadas por substantivos como "giraffo", "elefantina" e "flugzeug", em referência ao enredo do conto infantil, assim como por verbos básicos utilizados para a descrição de ações, como "kommen", por verbos auxiliares, como "haben" ou "werden", verbos modais, como "wollen" e, ainda, verbos que evidenciam a trama da narrativa, como "spielen".

Relativamente aos *hápax legomenon*, contabilizaram-se um total de 191 palavras, um número, ligeiramente inferior aos lemas com frequência igual a 1 presentes no subcorpus em PE (208 palavras). Esta diferença poderá, por um lado, indicar uma presença de menos erros ortográficos no AP, dada a elevada frequência de *hápax legomenon* baseados em *tokens* com erros

ortográficos. Atente-se na tabela 17, onde constam 50 entradas com os lemas com frequência igual a 1 no subcorpus em AP (consultar Anexo III).

Tabela 17

Lista de frequência com 50 entradas de hápax legomenon presentes no subcorpus em AP.

	Lemma	Frequency
1	and	1
2	abgefal	1
3	anfangte	1
4	anzuschimpfen	1
5	anlaufen	1
6	arm	1
7	beeindruckt	1
8	befürchten	1
9	begang	1
10	bei	1
11	benschbart	1
12	bekommen	1
13	bereit	1

	Lemma	Frequency
14	bermerkt	1
15	beschlossen	1
16	bewundern	1
17	bitten	1
18	brand	1
19	bringen	1
20	bösse	1
21	bössesie	1
22	chanstunns	1
23	dal	1
24	darum	1
25	das	1
26	der	1

	Lemma	Frequency
27	dengt	1
28	diese	1
29	dort	1
30	ebenfalls	1
31	ecke	1
32	eifersüchtig	1
33	einander	1
34	einfersüchtig	1
35	einige	1
36	einreden	1
37	ele-fantina	1
38	elefan-tina	1
39	elefant-ina	1

	Lemma	Frequency
40	elefanten-junge	1
41	elefantenjung	1
42	elefantinazu	1
43	elefantine	1
44	elefat	1
45	entfernt	1
46	entschuldigen	1
47	entwenden	1
48	erfolg	1
49	erfolgen	1
50	erklert	1

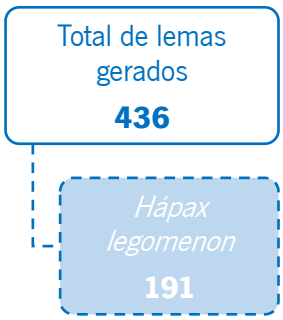
Para além da presença dos erros ortográficos, ressalta-se a presença de outro tipo de erros, por exemplo em 'wegnimmte' e 'bleibte'. Nestes dois casos percebe-se que o(s) informante(s) estão a aplicar as terminações de verbos regulares no *Präteritum* (pretérito perfeito) a verbos irregulares. A recorrência deste tipo de erros é bastante comum durante a aquisição de uma língua. Segundo Elsen (2000), "Researchers agree that children acquiring inflection show U-shaped behaviour curves. At first they use irregular as well as regular forms correctly. Then comes a period of overregularization and variation until the correct forms prevail" (p. 31). Montrul foi pioneira no trabalho sobre a aquisição de estruturas gramaticais em falantes de herança, observando, desta forma, várias semelhanças entre falantes de herança e falantes L2, notando que ambos os grupos partilham dificuldades com a morfologia flexional (Fernández-Dobao & Herschensohn, 2021, p. 2). Como assinala Montrul (2015), "(...) heritage speakers of diverse languages with overt morphological markings show strikingly similar patterns of omission of obligatory inflectional morphology and regularization of irregular forms, processes leading to overall simplification of inflectional morphology in their heritage languages" (p. 54).

Para além disso, muitos dos erros identificados no processo de lematização tiveram origem, por exemplo, na não utilização da inicial maiúscula em substantivos, como em "händ" (*Hand/ Hände*),

"freunde**n**" (*Freund/ Freunde*), "nähe**n**" (*Nähe*) e "wasser**n**" (*Wasser*). Nestes três últimos exemplos, o lematizador como não reconheceu que eram substantivos, por não estarem escritos com a inicial maiúscula, assumiu que seriam verbos, propondo, por isso, lemas com a terminação em "n" (terminação comum no infinitivo em AP). De igual forma, a concordância absoluta grafema-som, que induzem o informante em erro na produção escrita, foram identificados, por exemplo, nos lemas "spilsüg" (*Spielzeug*), "erklert" (*erklärt*) e "flugsöig" (*Flugzeug*). De destacar que o *token* "Flugzeug" (em português *avião*) foi um dos itens lexicais que mais variações ortográficas apresentou, de igual forma que o *token* "piscina" no subcorpus em PE. Um dos erros ortográficos comumente encontrados foi na omissão e/ou inclusão de grafemas, como em "kamm" (*kam*), "man" (*Mann*), "kapput" (*kaputt*), "geschaft" (*geschafft*), "bekamm" (*bekam*) e "herraus" (*heraus*). Por último, vários erros foram produzidos devido à criação de palavras compostas por parte do(s) informante(s), característica gramatical abundante no AP, dando origem, por exemplo, a lemas como "elefantendame", "elefanten-junge" e "elefantenmädchen". Em consequência, em vez de haver um só lema que agrupe as várias formas identificadas e derivadas, provavelmente, do nome "Elefantina", foram gerados vários lemas pelo lematizador. O diagrama da figura 11 esquematiza os resultados obtidos.

Figura 11

Representação do total de lemas e hápax legomenon gerados no subcorpus em AP.



Por sua vez, o *zu* foi um dos lemas com maior frequência encontrado no subcorpus em AP. Este item lexical é crucial na gramática alemã, particularmente em construções infinitivas, dado que desempenha um papel importante na estrutura das frases. A sua funcionalidade tanto pode servir para construções infinitivas, indicando propósito, necessidade ou intenção, como também pode servir de preposição para indicar movimento, direção, propósito ou adição (Schallert, 2020, p.

72). Na figura 12 podem observar-se algumas linhas de concordância que demonstram o comportamento deste item no subcorpus em AP (consultar Anexo IV).

Figura 12

Concordância do item "zu" no subcorpus em AP.

Linhas	Left context	KWIC	Right context
1	-<s>Giraffo hatte ein Spielzeug dabei, ein Flugzeug.</s><s>Er fing an damit	zu	spielen.</s><s>Elefantina schaute ihm beeindruckt zu.</s><s>Doch irgendv
2	><s>Er fing an damit zu spielen.</s><s>Elefantina schaute ihm beeindruckt	zu	.</s><s>Doch irgendwann wurde sie neidisch auf ihren Freund und nahm ih
3	war.</s><s>Doch irgendwann wurde er böse und fing an laut mit Elefantina	zu	schimpfen.</s><s>Aber da kam ein anderer Elefant der zu wisse schien was
4	ıut mit Elefantina zu schimpfen.</s><s>Aber da kam ein anderer Elefant der	zu	wisse schien was passiert war und wollte ihnen helfen.</s><s>Elefantina gir
5	sse schien was passiert war und wollte ihnen helfen.</s><s>Elefantina ging	zu	ihm und fragte ob er eine Idee hatte wie man das Flugzeug aus dem Wasse
6	das Flugzeug aus dem Wasser holen könne.</s><s>Giraffo schaute traurig	zum	Flugzeug.</s><s>Der Elefant lehnte sich nach vorn am Beckenrand und ver
7	r Elefant lehnte sich nach vorn am Beckenrand und versuchte das Flugzeug	zu	greifen.</s><s>Aber ohne Erfolg.</s><s>Er erklärte ihnen dass das Flugzeu
8	eifen.</s><s>Aber ohne Erfolg.</s><s>Er erklärte ihnen dass das Flugzeug	zu	weit weg schwimme.</s><s>Da fing Giraffo an zu weinen weil sein Flugzeuç
9	ihnen dass das Flugzeug zu weit weg schwimme.</s><s>Da fing Giraffo an	zu	weinen weil sein Flugzeug versinken würde Da kam aber eine clevere Elefar
10	Hand.</s><s>Sie fing an das Flugzeug mit dem Greifnetz aus dem Wasser	zu	fischen.</s><s>Als sie es aus dem Wasser gefischt hatte, gab sie es sofort (
11	wollte auch mit dem Flugzeug spielen.</s><s>Sie entschloss es ihm einfach	zu	entwenden und weiter zu spielen.</s><s>Elefantina war so unvorsichtig, das
12	eug spielen.</s><s>Sie entschloss es ihm einfach zu entwenden und weiter	zu	spielen.</s><s>Elefantina war so unvorsichtig, dass sie das Spielzeug ins W
13	<s>Er wollte helfen und versuchte mit aus gestreckten Armen das Flugzeug	zu	angeln.</s><s>Etwas dass ihm nicht gelang, da das Flugzeug schon zu wei
14	ɜug zu angeln.</s><s>Etwas dass ihm nicht gelang, da das Flugzeug schon	zu	weit weg vom Beckenrand war.</s><s>Giraffo wurde so traurig, dass er zu v
15	ı zu weit weg vom Beckenrand war.</s><s>Giraffo wurde so traurig, dass er	zu	weinen anfangte.</s><s>Eine weitere Elefantin, die die Situation beobachte
16	siht Elefantina das sein bester freund was in den Händen hatte und siht ihn	zu	wie Er mit den spielzeug spielte.</s><s>Und dann wurde sie irgen wann nei
17	><s>Und dann probierte der andere Elefant das spielz-eug aus dem wasser	zu	holen aber er kommte nicht rann.</s><s>Und dann sagte der an-dere Elefar

Como se pode observar pela figura 12, este item lexical é utilizado, maioritariamente, como preposição, indicando, como referido, direção ou movimento para a um local ou alvo. O uso frequente deste item como preposição, visível na linha 5 através do excerto "Elefantina ging **zu** ihm", reflete a utilização de contextos em que a informação espacial e/ou direcional é relevante no corpus. Este item é, também, usado para expressar propósito ou objetivo, revelando a presença de orações no corpus que exprimam a transmissão de intenções, objetivos ou razão por detrás de uma ação. Atente-se no excerto da linha 11, "Sie entschloss es ihm einfach **zu** entwenden". Por fim, o seu uso também é bastante comum na formulação de orações infinitivas. A sua abundância como preposição sugere um uso significativo de construções infinitivas no corpus, indicando ações ou estados que visam ou estão relacionados com um objetivo específico, por exemplo em "Er fing an damit **zu** spielen", na primeira linha. A utilização do *zu* como preposição pode sugerir que o corpus envolve níveis de proficiência linguística variados, uma vez que o uso adequado deste item lexical em contextos preposicionais requer uma boa compreensão da gramática alemã. Não

obstante, o seu uso no subcorpus em AP, é também frequente como advérbio, como é visível na tabela 18.

Tabela 18

Comportamento do item "zu" identificado no subcorpus em AP.

	Estrutura	Exemplo	Frequência
zu como preposição	objetos dativos de *zu*	*Elefantina rannte **zum** Elefant*	5
	objetos acusativos de *zu*	*Sie probiert mit den **zum** den flugzeug fangen*	1
	verbos com *zu*	*Elefantina kam sofort **zu** ihm gelaufen*	8
zu como advérbio	verbos modificados por *zu*	*Aber da kam ein anderer Elefant der **zu** wisse*	5

Os resultados apresentados do processo de lematização de ambos os subcorpora fornecem informações relevantes sobre os usos e padrões linguísticos, assim como dos desafios no processo de aquisição de estruturas gramaticais pelos informantes. Desta forma, a identificação e a análise de alguns dos tipos de erros contribuíram para uma compreensão mais aprofundada do processo de aquisição da língua. Com base nestes conhecimentos, a secção 5.5. analisa sucintamente os resultados do processo de etiquetação morfossintática do corpus. De ressalvar que os erros preservados nas narrativas e que, de certa forma, influenciaram o processo de lematização, vão ser, de igual forma, observados nos resultados obtidos através do processo de etiquetação morfossintática do corpus.

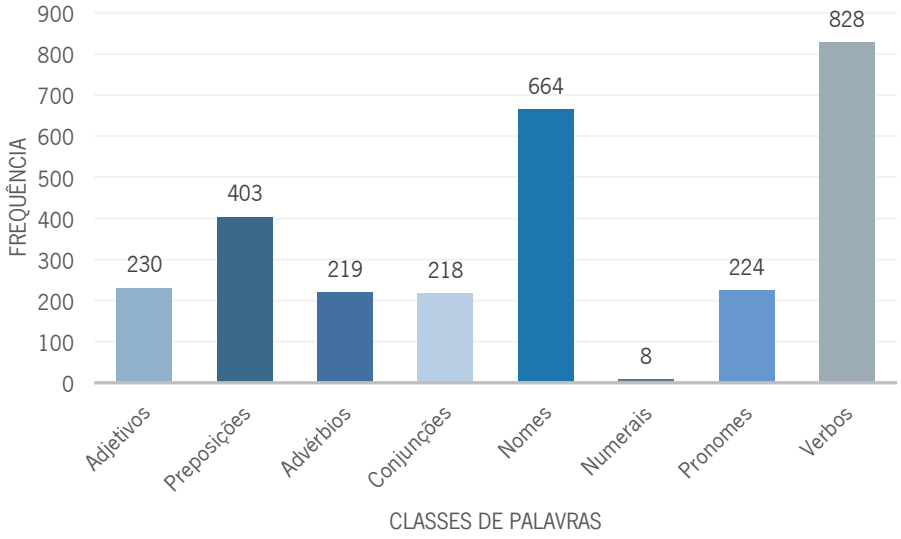
5.5. ETIQUETAÇÃO MORFOSSINTÁTICA DO SUBCORPUS EM PORTUGUÊS EUROPEU

Uma vantagem do *Sketch Engine* em comparação com outras ferramentas de gestão de corpora, é que a etiquetação morfossintática é feita a partir do momento em que o corpus é criado e compilado na plataforma, de acordo com o sistema gramatical da língua escolhida pelo utilizador aquando da inserção do corpus. Desta forma, este procedimento automático auxilia o investigador no processamento do corpus. Importa referir que as tarefas de PLN estão interligadas e que a natureza destas narrativas [erros ortográficos, morfossintáticos, etc.] induzem o anotador em erro. Como consequência, os resultados expressos no gráfico e tabelas abaixo representadas são produto da interpretação automática do anotador (pré-treinado).

A figura 13 mostra um gráfico de barras onde é possível observar resultado global do processo de anotação morfossintática, através do total de frequências em cada uma das principais classes de palavras no subcorpus em PE. Os **Verbos** foram a classe de palavras que mais ocorreu no subcorpus, contabilizando 161 itens lexicais e uma frequência de 828, sendo que *ficar*, *brincar*, *estar*, *tirar* e *ter* foram os cinco verbos mais utilizados pelos informantes, respetivamente. Seguem-se as classes de palavras dos **Nomes**, com 118 itens e uma frequência total de 664, as **Preposições** com 27 itens e uma frequência de 403, os **Adjetivos** com 37 itens que ocorreram 230 vezes, os **Pronomes** que registaram 16 itens e uma frequência total de 224, os **Advérbios** com 44 itens e 219 frequências totais, as **Conjunções** com apenas 7 itens e uma frequência de 218 e por último, os **Numerais**, onde foi identificado apenas 1 item lexical com uma frequência total de 8.

Figura 13

Gráfico representativo das várias classes de palavras presentes no subcorpus em PE e respetiva frequência.



5.5.1. Verbos

De acordo com Rühlemann (2014), as narrativas, independentemente de serem "grandes histórias" ou "pequenas histórias", contêm elementos no seu discurso, que se relacionam com a evidência linguística das histórias relatarem sequências de eventos que aconteceram numa situação remota da situação presente da narração da história (p. 317). Na maioria dos casos, esta

característica manifesta-se no uso de verbos no passado, itens lexicais que têm uma referência clara ao passado (esta manhã, ontem, etc.), referência a locais afastados do local de fala e referentes (tipicamente pessoas) não presentes na situação de narração (Rühlemann, 2014, p. 317). No subcorpus em análise, o recurso a verbos no passado é bastante marcado, pois a descrição sucessiva de acontecimentos neste género literário assim o exige. Através do processamento de anotação morfossintática automático, foram identificadas 28 diferentes classes de palavras, representadas na tabela 19. Desta forma, é possível observar as etiquetas correspondentes aos vários tempos e modos verbais identificados no subcorpus em PE, assim como o género e o número, seguidos da quantidade de vezes em que ocorreram nos textos (frequência) e um exemplo de uma palavra etiquetada retirada do subcorpus.

Tabela 19

Etiquetação morfossintática da classe de palavra "Verbos" no subcorpus em PE.

Posição	Etiqueta	Designação		Frequência	Exemplo
1	VMIS3S0	Pretérito perfeito simples	3.ª pess. sing.	323	*viu*
2	VMN0000	Infinitivo	-	219	*brincar*
3	VMII3S0	Pretérito imperfeito	3.ª pess. sing.	99	*estava*
4	VMP00SM	Particípio	Masc.	38	*trazido*
5	VMIP3S0	Presente do indicativo	3.ª pess. sing.	36	*tirou-le**
6	VMII1S0	Pretérito imperfeito	1.ª pess. sing.	26	*tinha*
7	VMIS3P0	Pretérito perfeito simples	3.ª pess. pl.	19	*começaram*
8	VMP00SF	Particípio	Fem.	12	*espantada*
9	VMII3P0	Pretérito imperfeito	3.ª pess. pl.	10	*olhavam*
10	VMIS1S0	Pretérito perfeito simples	1.ª pess. sing.	6	*consegui*
11	VMIP1S0	Presente do indicativo	1.ª pess. sing.	4	*consigo*
12	VMIM3P0	Pretérito mais-que-perfeito	3.ª pess. pl.	4	*encontraram-se*
13	VMIP2S0	Presente do indicativo	2.ª pess. sing.	4	*ajuda-los**
14	VMIP3P0	Presente do indicativo	3.ª pess. pl.	4	*são*
15	VMSP3S0	Presente do subjuntivo	3.ª pess. sing.	3	*contente*
16	VMSI3S0	Pretérito imperfeito do subjuntivo	3.ª pess. sing.	3	*abaijouse**
17	VMG0000	Gerúndio	-	3	*entre-gou-o**
18	VMIS2S0	Pretérito perfeito do indicativo	2.ª pess. sing.	2	*fizeste*
19	VMN03P0	Infinitivo	3.ª pess. pl.	2	*pensarem*
20	VMIF3P0	Futuro do indicativo	3.ª pess. pl.	2	*ficarão*
21	VMP00PM	Particípio passado	Pl. masc.	2	*animados*
22	VMIM1S0	Pretérito mais-que-perfeito	1.ª pess. sing.	1	*fora*
23	VMIP1P0	Presente do indicativo	1.ª pess. pl.	1	*pâs**
24	VMIC3S0	Condicional	3.ª pess. sing.	1	*poderia*
25	VMSP2S0	Presente do subjuntivo	2.ª pess. sing.	1	*ciumes**
26	VMSP3P0	Presente do subjuntivo	3.ª pess. pl.	1	*bambem**
27	VMN03S0	Infinitivo	3.ª pess. sing.	1	*esticar-se*
28	VMM02P0	Imperativo	2.ª pess. pl.	1	*consegui*

Na tabela 19, destaca-se, na primeira posição, as construções verbais no Pretérito Perfeito Simples, na 3.ª pessoa do singular, com uma frequência total de 323, seguido do Infinitivo, com uma frequência, também, bastante significativa, de 219. Na coluna "etiqueta" encontram-se as *POS-Tags* atribuídas às palavras do subcorpus, seguida da coluna "designação", onde explica o significado de cada etiqueta atribuída. Relativamente às etiquetas, estas são compostas por sete caracteres, que definem, no caso dos verbos, a categoria da palavra (se é verbo), o tipo (se é um verbo principal, auxiliar ou semiauxiliar), o modo, o tempo, a pessoa, o número e o género. Sempre que não existir um valor atribuído a um elemento, este é preenchido com um zero. Por exemplo, a etiqueta "VMIS3P0" (posição 7) corresponde a um verbo no pretérito perfeito simples na terceira pessoa do plural. Estas formas verbais ocorreram 19 vezes, como se observa na coluna, onde está designada a frequência.

Como mencionado, algumas etiquetas referentes a tempos verbais foram geradas com base em palavras que continham erros ortográficos, que são facilmente identificadas na tabela com o sinal de asterisco (*), como em *tirou-le*, *ajuda-los*, *abaijouse*, *entre-gou-o*, *pâs*, *ciumes* e *bambem*. De ressalvar que foram, também, identificadas palavras com erros ortográficos noutros tempos verbais, embora não apareçam designados na tabela 19. Em relação ao processo de anotação automático, que como já fora discutido, tem o objetivo de etiquetar o corpus de acordo com a definição e o contexto de cada palavra. No entanto, esse processo é, por vezes, falível. Atente-se no exemplo *consigo* (posição 11 da tabela 19), cujo tempo verbal atribuído foi o presente do indicativo. Note-se, agora, o contexto global através da frase completa retirada do subcorpus, que diz "O girafo tinha um brinquedo *consigo*". Neste caso, o anotador etiquetou a palavra *consigo* como se fosse um verbo, sem ter em consideração o contexto em que esta palavra se inseria. Desta forma, o etiquetador falhou na atribuição da etiqueta aquando do processo de desambiguação da palavra. Casos semelhantes foram, também, detetados no processo de anotação com as palavras *contente*[59] (posição 15), *animados*[60] (posição 21) e *fora*[61] (posição 22), cujas frases se encontram em nota de rodapé para uma melhor contextualização.

Por outro lado, identificou-se a presença da utilização de tempos verbais, como o uso do condicional, e construções frásicas elaboradas, que denotam um nível de proficiência linguístico

[59] "O Girafo esta muito *contente* ele e a Elefantina voltão a ser amigos".
[60] "Os dois estavam muito *animados* e contentes".
[61] "Fácilemente ela tirou o avião *fora* da água".

considerável. Atente-se nos exemplos *pensarem*[62] (posição 19) e *poderia*[63] (posição 24), que ocorreram apenas duas e uma vez, respetivamente.

5.5.2. Nomes

A classe de palavras dos nomes representa, em especial, a área temática das narrativas, dado que esta categoria gramatical representa 'coisas' ou áreas dentro de um domínio específico (Silva & Batoréo, 2010, p. 235). Com efeito, o *token* mais utilizado pelos informantes nas suas produções escritas foi *avião*, seguido de *Girafo* e *Elefante*, as duas personagens principais do enredo. A tabela 20 apresenta as 9 classes gramaticais identificadas pelo etiquetador no subcorpus em PE, com a respetiva designação e a sua frequência.

Tabela 20

Etiquetação morfossintática da classe de palavra "Nomes" no subcorpus em PE.

Posição	Etiqueta	Designação	Frequência	Exemplo
1	NCMS000	Nome comum, masc., sing.	280	*avião*
2	NCFS000	Nome comum, fem., sing.	162	*rede*
3	NPMS000	Nome próprio, masc., sing.	151	*Girafo*
4	NPFS000	Nome próprio, fem., sing.	22	*Senhora*
5	NCMP000	Nome comum, masc., pl.	21	*peixes*
6	NP00000	Nome próprio	17	*Girafino*
7	NCFP000	Nome comum, fem., pl.	8	*mãos*
8	NCCS000	Nome sobrecomum, sing.	2	*triste*
9	NPMP000	Nome próprio, masc., pl.	1	*Amigos*

Relativamente à classificação dos nomes, nota-se que a etiquetação variou entre nomes comuns e próprios, sendo diferenciada apenas pelo género e número das palavras. Desta forma, obtiveram-se várias combinações de etiquetas atribuídas aos nomes identificados no subcorpus em PE, constatando-se uma forte utilização por parte dos informantes de nomes comuns do género masculino e feminino no singular e nomes próprios do género masculino no singular, cujas categorias ocupam os três primeiros lugares na tabela 20.

Por outro lado, também se encontraram algumas irregularidades na etiquetação na classe dos Nomes. Na posição 8, a etiqueta com a designação "NCCS000", não corresponde à classe da

[62] "Depois de muito *pensarem*, chegou uma elefantina muito despachada que os queria ajudar".
[63] "Ele disse tudo isso mas não pensou que o avião *poderia* cair na água da piscina, porque foi exatamente isso que aconteceu".

palavra *triste*[64] de acordo com o contexto em que foi empregue. Também na posição 9 da tabela, encontra-se a etiqueta "NPMP000", cuja designação corresponde ao nome próprio masculino plural, onde a única ocorrência registada foi em *Amigos*. Provavelmente, o etiquetador atribuiu a etiqueta de nome próprio a esta palavra devido ao facto de esta estar com a inicial em maiúscula, embora não se encontre no início de uma frase. Denota-se, uma vez mais, a presença do fenómeno de transferência interlinguística a nível da ortografia, abordado na análise sobre a lematização do subcorpus em PE.

5.5.3. Preposições

As preposições são palavras invariáveis que conectam dois termos numa oração. Desta forma, o sentido do primeiro termo (antecedente) é explicado ou completado pelo segundo termo (consequente) (Cunha & Sintra, 2017, p. 569). Na tabela 21 é possível observar, as 7 categorias de preposições mais utilizadas pelos informantes na elaboração das narrativas em PE, a sua designação, respetiva frequência e exemplos.

Tabela 21

Etiquetação morfossintática da classe de palavra "Proposições" no subcorpus em PE.

Posição	Etiqueta	Designação	Frequência	Exemplo
1	SP	Preposição	289	*a*, *com*, *para*
2	SP+*	*[65]	80	*nesse*, *do*, *ao*
3	SP+DA	Preposição com adjetivo demonstrativo	27	*na*, *no*
4	SP+DI0FS0	Preposição com pronome indefinido, fem., sing.	4	*numa*
5	SP+PP3MP00	Preposição com pronome pessoal, 3.ª pess., masc., pl.	1	*deles*
6	SP+PD0NN00	Preposição com pronome demonstrativo, sing.	1	*disso*
7	SP+PP3MS00	Preposição com pronome pessoal, 3.ª pess., masc., sing.	1	*dele*

Observa-se que as preposições simples ('SP'), como *a*, *com*, *para*, *por*, *de*, *sem*, *até*, entre outras, foram as mais utilizadas pelos informantes na produção das suas narrativas em PE, denotando uma frequência de 289. Estas são seguidas de preposições com um complemento não especificado que, embora significativamente menos frequentes, constituem o segundo grupo de preposições mais utilizadas pelos informantes. Na terceira posição da tabela, encontram-se, também com menor expressão, preposições com adjetivos demonstrativos, que refletem a

[64] "Girafino *triste* comessou a chorar."

[65] Não foi possível encontrar uma designação precisa para esta etiqueta.

capacidade dos informantes em ligar as preposições a referências demonstrativas da história. Os restantes tipos de preposições identificadas pelo anotador representam casos excecionais de construções com menos frequência no subcorpus, mas que ao mesmo tempo remontam para uma utilização mais diversificada e concisa da língua por parte das crianças.

5.5.4. Adjetivos

Os adjetivos modificam os substantivos. Estes, como Cunha e Sintra (2017) apontam, servem, numa primeira instância para caracterizar os objetos, seres ou noções nomeadas pelo substantivo, atribuindo-lhe uma qualidade (ou defeito), modo de ser, aspeto ou aparência, estado e, em segundo lugar, para estabelecer uma relação com o substantivo em termos de tempo, espaço, matéria, finalidade, propriedade e procedência (p. 259). De acordo com a tabela 22, onde constam as seis classes de adjetivos identificadas no subcorpus, verifica-se uma utilização recorrente de adjetivos do género feminino e número singular ('AQ0FS00'). Importa frisar que a maioria das etiquetas "AQ0FS00" correspondem à palavra *Elefantina*, que, no conto, se referia a uma das personagens principais. Em PE, a palavra *elefantina* não tem significado, embora se perceba, de acordo com o contexto em que se insere, que a personagem é um elefante do sexo feminino. No entanto, o anotador, ao identificá-la, processou-a como um adjetivo.

Tabela 22

Etiquetação morfossintática da classe de palavra "Adjetivos" no subcorpus em PE.

Posição	Etiqueta	Designação	Exemplo	Frequência
1	AQ0FS00	Adjetivo qualificativo, fem., sing.	*Elefantina*, *desportiva*	131
2	AQ0CS00	Adjetivo qualificativo, comum, sing.	*triste*, *contente*, *feliz*	62
3	AQ0MS00	Adjetivo qualificativo, masc., sing.	*deishoulo**, *bravo*	21
4	AQ0CP00	Adjetivo qualificativo, comum, pl.	*felizes*, *contentes*, *tristes*	9
5	AQ0MP00	Adjetivo qualificativo, masc., pl.	*ajudasnos**, *juntos*	6
6	AQ0CN00	Adjetivo qualificativo, comum, invariável	*esticavass**	1

Observa-se, também, que a utilização dos adjetivos qualificativos incidiu, de forma geral, em palavras que descrevessem o estado de espírito das personagens, como *triste(s)*, *contente(s)*, *feliz(es)*, entre outras. Atente-se no exemplo da sexta posição, que dá conta de um adjetivo gerado com base num *token* com erro ortográfico. Através da funcionalidade *Concordance*, foi possível constatar que o informante, provavelmente, queria ter escrito "esticava-se" (do verbo *esticar-se*),

tendo escrito "esticavass"[66]. Desta forma, o etiquetador, identificou o *token* como um adjetivo, atribuindo-lhe a etiqueta de "AQ0CN00" (adjetivo qualificativo, género comum e número invariável). Este tipo de situações são recorrentes ao longo do corpus, devido à automatização da ferramenta, sendo, por isso, essencial uma análise minuciosa dos dados.

5.5.5. Pronomes

Os pronomes desempenham funções semelhantes às dos elementos nominais numa oração. Desta forma, podem representar ou acompanhar um substantivo, determinando-lhe a extensão do seu significado (Cunha & Sintra, 2017, p. 289). Com efeito, a utilização de pronomes no subcorpus em PE foi relativamente diversificada, tendo-se destacado a utilização dos pronomes pessoais *ele* e *ela*, ocupando as segunda e terceira posições na tabela 23, sendo que o mais utilizado pelos informantes foi o pronome relativo *que*, que ocupa a primeira posição.

Tabela 23

Etiquetação morfossintática da classe de palavra "Pronomes" no subcorpus em PE.

Posição	Etiqueta	Designação	Exemplos	Frequência
1	PR0CN00	Pronome relativo, comum, invariável	*que*	50
2	PP3MS00	Pronome pessoal, 3.ª pess., masc., sing.	*ele*	48
3	PP3FS00	Pronome pessoal, 3.ª pess., fem., sing.	*ela*	35
4	PP3CN00	Pronome pessoal, 3.ª pess., comum, invariável	*se*	18
5	PD0MS00	Pronome demonstrativo, masc., sing.	*o*	17
6	PD0NN00	Pronome demonstrativo, neutro, invariável	*isso, isto, aquilo*	11
7	PP3FSA0	Pronome pessoal, 3.ª pess., fem., sing. acusativo	*la**	7
8	PP3MP00	Pronome pessoal, 3.ª pess., masc., pl.	*eles*	7
9	PI0NN00	Pronome indefinido, neutro, invariável	*tudo*	6
10	PP3MSA0	Pronome pessoal, 3.ª pess., masc., sing., acusativo	*o*	6
11	PI0CN00	Pronome indefinido, comum, invariável	*algo*, *nada*	4
12	PP3MPA0	Pronome pessoal, 3.ª pess., masc., pl., acusativo	*os*	3
13	PI0MP00	Pronome indefinido, masc., pl.	*todos, outros*	3
14	PI0MS00	Pronome indefinido, masc., sing.	*outro, nenhum*	2
15	PI0FS00	Pronome indefinido, fem., sing.	*nenhuma, toda*	2
16	PP3CSD0	Pronome pessoal, 3.ª pess., comum, sing., dativo	*lhe*	1
17	PP3FPA0	Pronome pessoal, 3.ª pess., fem., pl., acusativo	*as*	1
18	PD0FS00	Pronome demonstrativo, fem., sing.	*esta**	1
19	PP3CPD0	Pronome pessoal, 3.ª pess., comum, pl., dativo	*lhes*	1
20	PP3CSO0	Pronome pessoal, 3.ª pess., comum, sing.	*consigo*	1

[66] "(...) O elefante tentou ajudar *esticavass* e esticavasse mas não consegui (...)."

De destacar a quase nula existência de *tokens* com erros nesta classe de palavras, com exceção das palavras *la* (posição 7) e *esta* (posição 18). Neste último caso, embora a palavra *esta* corresponda, efetivamente, a um pronome demonstrativo feminino no singular, nota-se que o informante se esqueceu de colocar o acento no "a", dado que o contexto em que a palavra se insere no subcorpus indica que este se referia à forma "está" no presente do indicativo do verbo *estar*[67]. Como consequência, a falta de acentuação levou a que o anotador etiquetasse a palavra como um pronome demonstrativo. Para além disso, importa referir a presença de pronomes pessoais no acusativo (posições 10 e 12), em "Quando ela *o* alcançou (...)" ou "(...), porque a Elefantina *o* podia deixar cair.", que denotam a utilização de estruturas sintáticas mais complexas, revelando, desta forma, um certo domínio linguístico por parte do(s) informante(s).

5.5.6. Advérbios

Para além de adicionar ou complementar o contexto no discurso, os advérbios têm a particularidade de funcionarem como adjacentes de adjetivos, de advérbios e até de nomes (Marçalo, 2009, pp. 106-107). Relativamente ao uso dos advérbios por parte dos informantes, foram identificadas duas grandes categorias pelo etiquetador morfossintático, designadamente advérbios (*general adverbs*) e advérbios de negação (*adverbs of negation*), como mostra a tabela 24. Estes elementos linguísticos constituem uma componente vital da linguagem e da comunicação, uma vez que possuem a capacidade única de modificar verbos, adjetivos ou outros advérbios, transmitindo informações cruciais sobre tempo, lugar, modo, grau ou frequência (Cunha & Sintra, 2017, p. 555).

Tabela 24

Etiquetação morfossintática da classe de palavra "Advérbios" no subcorpus em PE.

Posição	Etiqueta	Designação	Exemplos	Frequência
1	RG	Advérbio	*muito, mais, também*	193
2	RN	Advérbio de negação	*não*	26

[67] "O Girafo *esta* muito contente ele e a Elefantina voltão a ser amigos."

De acordo com a tabela 24, os advérbios gerais foram, claramente, os mais utilizados na produção das narrativas em PE, sendo que, por oposição, o *token* "não" foi a única forma adverbial utilizada pelos informantes para expressar a negação, tendo sido empregue 26 vezes no subcorpus em PE.

5.5.7. Conjunções

À semelhança dos advérbios, também as conjunções presentes no subcorpus em PE foram categorizadas em dois grandes grupos – coordenativas e subordinativas. Estas desempenham um papel importante na estruturação de frases e no estabelecimento de relações significativas entre ideias, permitindo a articulação de pensamentos, a descrição de relações de causa e efeito, bem como a transmissão de contrastes (Cunha & Sintra, 2017, p. 593). Compreender os diversos tipos de conjunções é fundamental para perceber como é que a linguagem atinge a coerência e a coesão no discurso. A tabela 25 mostra a frequência destas no subcorpus em PE, sendo que as conjunções coordenativas foram as mais usadas na produção das narrativas. Dado que as conjunções coordenativas constituem um dos elementos linguísticos mais básicos numa língua para a conexão de orações, compreende-se a forte utilização por parte dos informantes.

Tabela 25

Etiquetação morfossintática da classe de palavra "Conjunções" no subcorpus em PE.

Posição	Etiqueta	Designação	Exemplos	Frequência
1	CC	Conjunção coordenativa	*e, mas*	164
2	CS	Conjunção subordinativa	*que, se, pois*	54

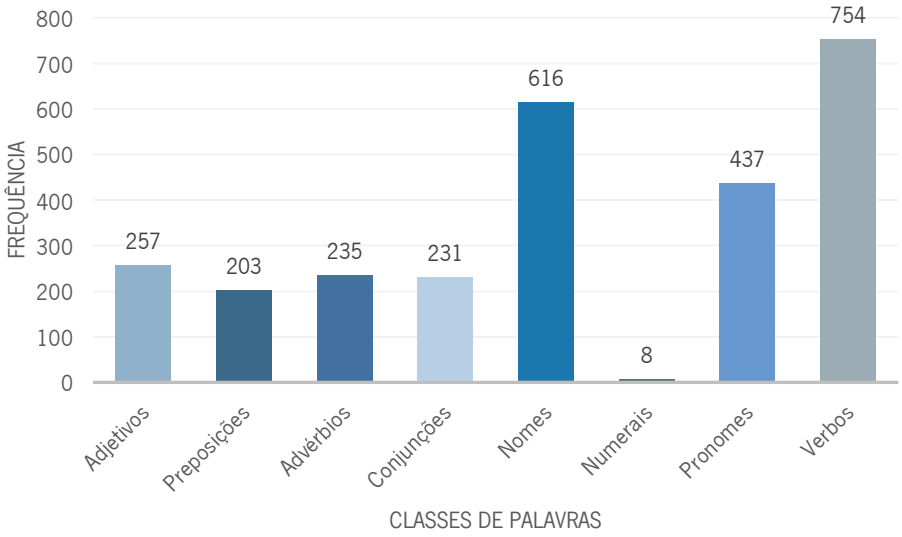
Relativamente à utilização de conjunções subordinativas, observa-se uma menor utilização por parte dos informantes nas suas narrativas, visto que constituem um certo grau de complexidade sintática e a sua utilização ocorre mais tardiamente em comparação às conjunções coordenativas, como, aliás, Glória *et al.* (2016) referem no seu estudo. De ressalvar que a conjunção subordinativa *pois* em "Chegou um elefante que os queria ajudar, *pois* tinha acompanhado a situação", foi uma das que teve menor frequência, tendo sido utilizada apenas uma única vez, por um dos informantes.

5.6. ETIQUETAÇÃO MORFOSSINTÁTICA DO CORPUS EM ALEMÃO PADRÃO

No subcorpus em AP, a diferença dos resultados ao nível dos itens e frequência das várias classes em relação aos resultados obtidos na etiquetação morfossintática do subcorpus em PE é pouco significativa. Atente-se no gráfico da figura 14, onde se expressam ao nível da frequência as várias classes de palavras etiquetadas no subcorpus em AP.

Figura 14

Gráfico representativo das várias classes de palavras presentes no subcorpus em AP e respetiva frequência.



Como é possível observar, os **Verbos** contabilizam 139 itens e uma frequência total de 754, sendo que *sein*, *haben*, *spielen*, *wollen* e *kommen* ocupam as cinco primeiras posições na tabela dos verbos mais usados pelos informantes. Seguem-se as classes dos **Nomes** com 88 itens e uma frequência de 616, os **Adjetivos** com 101 itens e uma frequência total de 257, os **Advérbios** com 52 itens e uma frequência de 235 e as **Conjunções**, onde foram identificados 17 itens lexicais e uma frequência total de 231. Por último, a classe de palavras dos **Numerais** contém 4 itens lexicais e uma frequência total igual ao subcorpus em PE, que foi de 8. Foram, ainda, identificadas várias classes de palavras que contêm itens lexicais com erros ortográficos, nomeadamente nos adjetivos, preposições, advérbios, conjunções, nomes e verbos.

Notou-se, no entanto, uma maior utilização de preposições por parte dos informantes na elaboração das narrativas em PE, quando comparados os resultados do subcorpus em AP. A classe das **Preposições** do subcorpus em AP teve uma frequência de 203, 200 a menos que no subcorpus em PE. A tabela 26 apresenta os níveis de frequência das várias classes de palavras resultantes do processamento automático de etiquetação morfossintática nos dois subcorpora. Por outro lado, o reverso aconteceu com a classe de palavras dos **Pronomes**, onde se registou uma maior frequência no subcorpus em AP, 213 itens a mais que no subcorpus em PE.

Tabela 26

Comparação dos resultados obtidos no processamento de etiquetação morfossintático dos subcorpora em PE e AP. Os resultados apresentam o nível de ocorrências (frequência) em cada classe de palavras.

	Subcorpus em PE	Subcorpus em AP
VERBOS	828 (+74)	754
ADJETIVOS	230	257 (+27)
PREPOSIÇÕES	**403 (+200)**	**203**
ADVÉRBIOS	219	235 (+16)
CONJUNÇÕES	218	231 (+13)
NOMES	664 (+48)	616
PRONOMES	**224**	**437 (+213)**
NUMERAIS	8	8

Com esta informação, pretende-se apenas apresentar e constatar factos com base nos resultados obtidos, não havendo, por isso, meios para indagar sobre o porquê da diferença acentuada entre a utilização de preposições e pronomes nas duas línguas. Não obstante, uma análise pormenorizada destes resultados poderá ser relevante para futuras investigações.

5.6.1. Verbos

A categoria de verbos mais utilizada pelos informantes na redação das suas narrativas em AP foram os verbos na terceira pessoa do pretérito perfeito do indicativo, como em "Er *fing an* damit zu spielen". Seguiram-se os verbos auxiliares também na terceira pessoa do pretérito perfeito do indicativo, como em "Als sie es aus dem Wasser gefischt *hatte*, gab sie es sofort Giraffo" e as construções com recurso ao infinitivo, como por exemplo em "Da fing Giraffo an zu *weinen* (...)". A tabela 27 mostra as 26 categorias verbais usadas nas narrativas em AP.

Tabela 27

Etiquetação morfossintática da classe de palavra "Verbos" no subcorpus em AP.

Posição	Etiqueta	Designação	Exemplo	Frequência
1	VFIN.Full.3.Sg.Past.Ind	Pretérito perfeito do indicativo, 3.ª pess., sing.	*fing*	256
2	VFIN.Aux.3.Sg.Past.Ind	Verbo auxiliar no pretérito perfeito do indicativo, 3.ª pess., sing.	*hatte*	124
3	VINF.Full	Infinitivo	*helfen*	117
4	VPP.Full.Psp	Particípio passado	*passiert*	44
5	VFIN.Aux.3.Sg.Pres.Ind	Verbo auxiliar no presente do indicativo, 3.ª pess., sing.	*ist*	40
6	VFIN.Full.3.Sg.Pres.Ind	Presente do indicativo, 3.ª pess., sing.	*kommt*	38
7	VFIN.Mod.3.Sg.Past.Ind	Verbo modal no pretérito perfeito do indicativo, 3.ª pess., sing.	*wollte*	35
8	VFIN.Full.3.Pl.Past.Ind	Pretérito perfeito do indicativo, 3.ª pess., pl.	*spielten*	21
9	VIMP.Full.2.Sg	Imperativo, 2.ª pess., sing.	*wasser**	18
10	VFIN.Aux.3.Pl.Past.Ind	Verbo auxiliar no pretérito perfeito do indicativo, 3.ª pess., pl.	*waren*	11
11	VFIN.Full.3.Sg.Pres.Subj	Presente do subjuntivo, 3.ª pess., sing.	*nähe**	8
12	VFIN.Aux.3.Sg.Past.Subj	Verbo auxiliar no pretérito perfeito do subjuntivo, 3.ª pess., sing.	*würde*	6
13	VFIN.Aux.3.Sg.Pres.Subj	Verbo auxiliar no presente do subjuntivo, 3.ª pess., sing.	*sei*	5
14	VFIN.Aux.3.Pl.Pres.Ind	Verbo auxiliar no presente do indicativo, 3.ª pess., pl.	*sind/haben*	4
15	VFIN.Full.3.Pl.Pres.Ind	Presente do indicativo, 3.ª pess., pl.	*traffen**	4
16	VINF.Aux	Auxiliar no infinitivo	*haben*	4
17	VFIN.Mod.3.Sg.Past.Subj	Verbo modal no pretérito perfeito do subjuntivo, 3.ª pess., sing.	*könnte*	3
18	VFIN.Aux.2.Sg.Pres.Ind	Verbo auxiliar no presente do indicativo, 2.ª pess., sing.	*hast*	3
19	VIMP.Full.2.Pl	Imperativo, 2.ª pess., pl.	*nehmt*	2
20	VFIN.Mod.3.Sg.Pres.Subj	Verbo modal no presente do subjuntivo, 3.ª pess., sing.	*könne**	2
21	VFIN.Mod.3.Sg.Pres.Ind	Verbo modal no presente do indicativo, 3.ª pess., sing.	*kann*	2
22	VFIN.Full.1.Pl.Past.Ind	Pretérito perfeito do indicativo, 1.ª pess., pl.	*spielten**	2
23	VINF.Full.zu	Infinitivo com partícula "zu"	*anzuschimpfen*	2
24	VFIN.Full.2.Pl.Pres.Ind	Presente do indicativo, 2.ª pess., pl.	*nehmt*	1
25	VFIN.Full.1.Sg.Pres.Ind	Presente do indicativo, 1.ªpess., sing.	*urhein**	1
26	VFIN.Mod.3.Pl.Past.Ind	Verbo modal no pretérito perfeito do indicativo, 3.ª pess., pl.	*wollten*	1

Relativamente às etiquetas, estas são compostas por seis elementos, separados por um ponto final. Atente-se no exemplo "VFIN.Full.3.Sg.Past.Ind". O primeiro elemento ('VFIN') indica, neste caso, a categoria do verbo (verbo finito), o segundo elemento ('Full') indica o tipo de verbo (se é principal, auxiliar ou modal), o terceiro elemento ('3') indica a pessoa, o quarto elemento ('Sg') indica o número, o quinto elemento ('Past') indica o tempo e o sexto elemento ('Ind') indica o modo do verbo. Quando alguns dos elementos não são preenchidos por falta de informação, é colocado um asterisco, como por exemplo em "N.Name.Nom.Sg.*" (ver tabela 28).

Foram identificadas algumas lacunas no processo de etiquetação morfossintático, nomeadamente na posição 9 e 10 da tabela. A maioria dos verbos etiquetados automaticamente com a categoria "VIMP.Full.2.Sg" (verbo imperativo, segunda pessoa do singular), correspondem a substantivos que os informantes não escreveram com a inicial maiúscula. Alguns exemplos dos *tokens* etiquetados são "wasser", "netz", "freunde", "spiel" e "erfolg".

Para além disso, nem todas as ocorrências do verbo *sein* na figura 15 correspondem à função de verbo auxiliar. Veja-se, por exemplo, as ocorrências de *sein* nas linhas 7 e 9, onde este é utilizado como verbo principal.

Figura 15

Concordância dos verbos etiquetado com a etiqueta "VFIN.Aux.3.Pl.Past.Ind".

Linhas	Left context	KWIC	Right context
1	angeln. sie hat es geschaft und gab das Flugzeug an Giraffo.</s><s>Beide	waren	wieder froh Giraffo war froh weil er sein Flugzeug hatte und Elefantina war fr
2	angeln. sie hat es geschaft und gab das Flugzeug an Giraffo.</s><s>Beide	waren	wieder froh Giraffo war froh weil er sein Flugzeug hatte und Elefantina war fr
3	<s>Sie hatte das geschaffen.</s><s>Giraffo war wieder fröhlich.</s><s>Sie	waren	wieder gute Freunden mit einander.</s><s>Eines Tages sieth Elefantina sei
4	ɜg. Sie gab den Spielzeug Girafo, er war sehr glücklich.</s><s>Zum schluss	waren	Girafo und Elefantina wieder freunde.</s><s>Eines Tages waren zwei Freun
5	ım schluss waren Girafo und Elefantina wieder freunde.</s><s>Eines Tages	waren	zwei Freunden zusamen Elefantina und Giraffo, Sie waren neben ein pool.</
6	><s>Eines Tages waren zwei Freunden zusamen Elefantina und Giraffo, Sie	waren	neben ein pool.</s><s>Giraffo hatte ein Flugzeug, Elefantina war gespanen
7	as Spielzeug.</s><s>Er nahm es überglücklich wieder in die Hand Und nun	waren	Elefantina und Giraffo wieder gute Freunde.</s><s>Eines Tages traffen sich
8	sfischen und gab es Giraffe zurück Elefantiene entschuldigte sich und beide	wurden	wieder beste Freunde.</s><s>Eines Tages gingen ein verspielter Giraffen Ju
9	ɜie es hatte gab sie das Spielzeug Giraffo der überglücklich war.</s><s>Alle	waren	wieder glücklich: Giraffo hatte seinen Spielzeug wieder und Elefantina sah d
10	ihr Freund wieder glücklich war.</s><s>Eines Tages, Elefantina und Girraffo	hatten	sich getroffen.</s><s>Giraffo hatte ein brand neues "Spielflugzeug" dabei!</
11	var bereit Ihnen zu helfen.</s><s>Während Sie es versuchte raus zu picken	waren	alle 3 sehr froh!</s><s>Ja!</s><s>Die Frau hats geschafft und sie überreich

Por último, atente-se nos exemplos das posições 19 e 24 da tabela 27, cujo verbo conjugado "nehmt" foi etiquetado de duas maneiras diferentes, de acordo com o contexto em que foi utilizado pelos informantes. Na posição 19 observa-se que a etiqueta atribuída corresponde à forma do Imperativo na 2.ª pessoa do plural ('VIMP.Full.2.Pl') e na posição 24, a etiqueta atribuída corresponde à forma do Presente do indicativo na 2.ª pessoa do plural ('VFIN.Full.2.Pl.Pres.Ind'). Desta forma, é percetível, como já fora referido, o papel desambiguador que o etiquetador morfossintático tem durante o processamento do corpus.

5.6.2. Nomes

Ao analisar os *tokens* que foram etiquetados como **Nomes**, a primeira posição é ocupada maioritariamente pelos *tokens* "Giraffo" e "Elefantina". Estes são derivados dos nomes comuns em alemão *Giraffe* (girafa) e *Elefant* (elefante), que no contexto das narrativas correspondem aos nomes próprios das personagens da história. Desta forma, é natural que estes ocorram com bastante frequência no corpus. A tabela 28 mostra as categorias de nomes identificados no subcorpus em AP, verificando-se uma forte utilização de nomes comuns por parte dos informantes nas suas produções escritas.

Tabela 28

Etiquetação morfossintática da classe de palavra "Nomes" no subcorpus em AP.

Posição	Etiqueta	Designação	Exemplo	Frequência
1	N.Name.Nom.Sg.*	Nome próprio, nominativo, sing.	*Giraffo/Elefantina*	146
2	N.Reg.Acc.Sg.Neut	Nome comum, acusativo, sing., neutro	[ein] *Spielzeug*	112
3	N.Reg.Dat.Sg.Neut	Nome comum, dativo, sing., neutro	[aus dem] *Wasser*	54
4	N.Reg.Nom.Sg.Masc	Nome comum, nominativo, sing., masc.	*Elefant*	46
5	N.Reg.Nom.Sg.Neut	Nome comum, nominativo, sing., neutro	*Flugzeug*	35
6	N.Reg.Nom.Sg.Fem	Nome comum, nominativo, sing., fem.	*Elefantin*(e)*	28
7	N.Reg.Gen.Sg.Masc	Nome comum, genitivo, sing., masc.	*Tages*	21
8	N.Name.Nom.Sg.Masc	Nome próprio, nominativo, sing., masc.	*Giraffo/Elefantina(?)*	19
9	N.Reg.Dat.Sg.Fem	Nome comum, dativo, sing., fem.	[in der] *Hand*	19
10	N.Reg.Acc.Sg.Masc	Nome comum, acusativo, sing., masc.	*Erfolg*	17
11	N.Name.Dat.Sg.*	Nome próprios, dativo, sing.	[von] *Giraffo*	16
12	N.Name.Dat.Sg.Neut	Nome próprio, dativo, sing., neutro	[von] *Elefantina*	16
13	N.Reg.Acc.Sg.Fem	Nome comum, acusativo, sing., fem.	*Idee*	14
14	N.Name.Nom.Sg.Neut	Nome próprio, nominativo, sing., neutro	*Giraffo*	12
15	N.Name.*.*.*	Nome próprio	*dan**	10
16	N.Reg.Dat.Sg.Masc	Nome comum, dativo, sing., masc.	[zum] *Elefant*	9
17	N.Name.Acc.Sg.*	Nome próprio, acusativo, sing.	*Elefantina*	8
18	N.Name.Acc.Sg.Neut	Nome próprio, acusativo, sing., neutro	[für] *Giraffo*	8
19	N.Reg.Acc.Pl.Masc	Nome comum, acusativo, pl., masc.	*Freunde*	3
20	N.Name.Nom.Sg.Fem	Nome próprio, nominativo, sing., fem.	*Elefantina*	3
21	N.Reg.Dat.Pl.Masc	Nome comum, dativo, pl., masc.	*Freunden*	3
22	N.Reg.Nom.Pl.Masc	Nome comum, nominativo, pl., masc.	*Freunde*	3
23	N.Reg.Acc.Pl.Neut	Nome comum, acusativo, pl., neutro	*Kinder*	2
24	N.Reg.Dat.Pl.Fem	Nome comum, dativo, pl., fem.	*Händen*	2
25	N.Name.Dat.Sg.Fem	Nome próprio, dativo, sing., fem.	*Badi**	2
26	N.Reg.Nom.Pl.Fem	Nome comum, nominativo, pl., fem.	*Giraffen*	2
27	N.Reg.Gen.Sg.Fem	Nome comum, genitivo, sing., fem.	*Badi**	1
28	N.Name.Gen.Sg.*	Nome próprio, genitivo, sing.	*Giraffos*	1
29	N.Reg.Gen.Sg.Neut	Nome comum, genitivo, sing., neutro	*Schwimmbads*	1
30	N.Reg.Acc.Pl.Fem	Nome comum, acusativo, pl., fem.	*Vorkenntnisse*	1
31	N.Name.Gen.Sg.Neut	Nome próprio, genitivo, sing., neutro	*hats**	1

Observa-se, também, que as palavras de *Elefantina* e *Giraffo*, foram etiquetadas de diferentes maneiras durante o processo automático de anotação, correspondentes aos contextos em que apareciam. Alguns exemplos são visíveis nas posições 8, 11, 12 e 14 da tabela 28.

Para uma exploração dos resultados através da ferramenta *Sketch Engine*, procedeu-se, ainda, ao manuseamento das funcionalidades disponibilizadas, a fim de observar o comportamento dos informantes na redação das suas narrativas em AP. Desta forma, com a adição de filtros de pesquisa que incidiam sobre os casos gramaticais (nominativo, acusativo, dativo e genitivo), conseguiu-se averiguar uma elevada frequência de nomes próprios no nominativo (frequência de 180), seguido de nomes comuns no acusativo (frequência de 149) e na terceira posição, com uma frequência de 115, de nomes comuns no nominativo. Estes resultados não se encontram

expressos em nenhuma tabela nesta dissertação, uma vez que foram obtidos através do processamento automático de etiquetação morfossintática, que já se encontra exposto na tabela 28. Quanto à combinação [caso gramatical+número+género], a primeira posição, com uma frequência de 146, é ocupada por nomes no nominativo, singular e sem género identificado. Na segunda posição observa-se a utilização de nomes no acusativo, singular e género neutro, com uma frequência de 120 e na terceira posição, com uma frequência de 70, encontram-se os nomes no dativo, singular e género neutro. Percebe-se, também, que houve uma elevada frequência de nomes no singular em comparação com os nomes no plural.

5.6.3. Pronomes

Relativamente ao uso de pronomes, constatam-se, através da tabela 29, as categorias de pronomes identificadas pelo etiquetador morfossintático, uma utilização generalizada de pronomes pessoais na elaboração das narrativas em AP, que tiveram uma frequência de 291. Pese embora uma diferença considerável, os informantes recorreram, também, ao uso de pronomes possessivos (frequência de 43), indefinidos (frequência de 42), demonstrativos (frequência de 24), reflexivos (frequência de 21), relativos (frequência de 12) e interrogativos (frequência de 4).

Tabela 29

Etiquetação morfossintática da classe de palavra "Pronomes" no subcorpus em AP.

Posição	Etiqueta	Designação	Exemplo	Frequência
1	PRO.Pers.Subst.3.Nom.Sg.Masc	Pronome pessoal substituto, 3.ª pess., nom., sing., masc.	*er*	81
2	PRO.Pers.Subst.3.Nom.Sg.Fem	Pronome pessoal substituto, 3.ª pess., nom., sing., fem.	*sie*	80
3	PRO.Pers.Subst.3.Acc.Sg.Neut	Pronome pessoal substituto, 3.ª pess., acus., sing., neutro	*es*	28
4	PRO.Refl.Subst.3.Acc.Sg.*	Pronome reflexivo substituto, 3.ª pess., acus., sing.	*sich*	17
5	PRO.Pers.Subst.3.Nom.Sg.Neut	Pronome pessoal substituto, 3.ª pess., nom., sing., neutro	*es*	16
6	PRO.Pers.Subst.3.Nom.Pl.*	Pronome pessoal substituto, 3.ª pess., nom., pl.	*sie*	15
7	PRO.Pers.Subst.3.Dat.Sg.Masc	Pronome pessoal substituto, 3.ª pess., dativo, sing., masc.	*ihm*	14
8	PRO.Pers.Subst.3.Acc.Pl.*	Pronome pessoal substituto, 3.ª pess., acus., pl.	*sie*	14
9	PRO.Dem.Subst.Acc.Sg.Neut	Pronome demonstrativo substituto, acus., sing., neutro	*das*	13
10	PRO.Poss.Attr.Acc.Sg.Neut	Pronome possessivo atributivo, acus., sing., neutro	*sein/mein/ihr*	13
11	PRO.Pers.Subst.3.Acc.Sg.Masc	Pronome pessoal substituto, 3.ª pess., acus., sing., masc.	*ihn*	12
12	PRO.Pers.Subst.3.Dat.Pl.*	Pronome pessoal substituto, 3.ª pess., dativo, pl.	*ihnen*	12
13	PRO.Indef.Subst.*.*.Neut	Pronome indefinido substituto, neutro	*was/nichts (1)*	9
14	PRO.Indef.Subst.Nom.Pl.*	Pronome indefinido substituto, nom., pl.	*alle*	9
15	PRO.Poss.Attr.Acc.Sg.Masc	Pronome possessivo atributivo, acus., sing., masc.	*ihren*	7
16	PRO.Dem.Subst.Nom.Sg.Neut	Pronome demonstrativo substituto, nom., sing., neutro	*das*	6
17	PRO.Poss.Attr.Acc.Sg.Fem	Pronome possessivo atributivo, acus., sing., fem.	*seine*	5

18	PRO.Poss.Attr.Dat.Sg.Neut	Pronome possessivo atributivo, dativo, sing., neutro	*seinem*	5
19	PRO.Pers.Subst.3.Dat.Sg.Fem	Pronome pessoal substituto, 3.ª pess., dativo, sing., fem.	*ihr*	5
20	PRO.Poss.Attr.Nom.Sg.Neut	Pronome possessivo atributivo, nom., sing., neutro	*sein*	4
21	PRO.Inter.Subst.Nom.Sg.Neut	Pronome interrogativo substituto, nom., sing., neutro	*was*	4
22	PRO.Indef.Subst.Dat.Sg.Fem	Pronome indefinido substituto, dativo, sing., fem.	*einen*	4
23	PRO.Indef.Subst.Nom.Sg.*	Pronome indefinido substituto, nom., sing.	*man*	4
24	PRO.Pers.Subst.3.Acc.Sg.Fem	Pronome pessoal substituto, 3.ª pess., acus., sing., fem.	*sie*	4
25	PRO.Indef.Subst.Nom.Sg.Neut	Pronome indefinido substituto, nom., sing., neutro	*anderes*	4
26	PRO.Pers.Subst.2.Nom.Sg.*	Pronome pessoal substituto, 2.ª pess., nom., sing.	*du*	3
27	PRO.Poss.Attr.Nom.Sg.Masc	Pronome possessivo atributivo, nom., sing., masc.	*ihr/sein (1)*	3
28	PRO.Dem.Subst.Dat.Sg.Neut	Pronome demonstrativo substituto, dativo, sing., neutro	*dem*	3
29	PRO.Pers.Subst.1.Dat.Sg.*	Pronome pessoal substituto, 1.ª pess., dativo, sing.	*mir*	3
30	PRO.Refl.Subst.3.Acc.Pl.*	Pronome reflexivo substituto, 3.ª pess., acus., pl.	*sich*	3
31	PRO.Rel.Subst.Nom.Sg.Masc	Pronome relativo substituto, nom., sing., masc.	*der*	3
32	PRO.Rel.Subst.Nom.Sg.Fem	Pronome relativo substituto, nom., sing., fem.	*die*	2
33	PRO.Rel.Subst.Dat.Sg.Neut	Pronome relativo substituto, dativo, sing., neutro	*dem*	2
34	PRO.Indef.Subst.Acc.Sg.Neut	Pronome indefinido substituto, acus., sing., neutro	*alles/was*	2
35	PRO.Indef.Attr.*.*.Neut	Pronome indefinido atributivo, neutro	*nichts*	2
36	PRO.Indef.Attr.Dat.Sg.Fem	Pronome indefinido atributivo, dativo, sing., fem.	*einiger/aller*	2
37	PRO.Poss.Attr.Dat.Pl.Fem	Pronome possessivo atributivo, dativo, pl., fem.	*ihren/seinen*	2
38	PRO.Pers.Subst.1.Nom.Pl.*	Pronome pessoal substituto, 1.ª pess., nom., pl.	*wir*	2
39	PRO.Indef.Subst.*.*.*	Pronome indefinido substituto	*mehr*	2
40	PRO.Poss.Attr.Dat.Sg.Masc	Pronome possessivo atributivo, dativo, sing., masc.	*ihrem*	1
41	PRO.Dem.Subst.Nom.Sg.Fem	Pronome demonstrativo substituto, nom., sing., fem.	*diese*	1
42	PRO.Indef.Subst.Acc.Sg.Masc	Pronome indefinido substituo, acus., sing., masc.	*einen*	1
43	PRO.Indef.Attr.*.*.*	Pronome indefinido atributivo	*viel*	1
44	PRO.Indef.Attr.Nom.Sg.Neut	Pronome indefinido atributivo, nom., sing., neutro	*alles*	1
45	PRO.Rel.Subst.Acc.Sg.Masc	Pronome relativo substituto, acus., sing., masc.	*den*	1
46	PRO.Rel.Subst.Nom.Sg.Neut	Pronome relativo substituto, nom., sing., neutro	*was*	1
47	PRO.Poss.Attr.Dat.Pl.Neut	Pronome possessivo atributivo, dativo, pl., neutro	*seinen*	1
48	PRO.Dem.Subst.Dat.Sg.Fem	Pronome demonstrativo substituto, dativo, sing., fem.	*der*	1
49	PRO.Pers.Subst.3.Dat.Sg.*	Pronome pessoal substituto, 3.ª pess., dativo, sing.	*ihm*	1
50	PRO.Rel.Subst.Dat.Sg.Fem	Pronome relativo substituto, dativo, sing., fem.	*der*	1
51	PRO.Refl.Subst.3.Dat.Pl.*	Pronome reflexivo substituto, 3.ª pess., dativo, pl.	*einander*	1
52	PRO.Rel.Subst.Acc.Sg.Neut	Pronome relativo substituto, acus., sing., neutro	*das*	1
53	PRO.Rel.Subst.Dat.Sg.Masc	Pronome relativo substituto, dativo, sing., masc.	*dem*	1
54	PRO.Pers.Subst.3.Dat.Sg.Neut	Pronome pessoal substituto, 3.ª pess., dativo, sing., neutro	*ihm*	1
55	PRO.Poss.Attr.Gen.Sg.Masc	Pronome possessivo atributivo, genitivo, sing., masc.	*ihres*	1
56	PRO.Poss.Attr.Dat.Sg.Fem	Pronome possessivo atributivo, dativo, sing., fem.	*seiner*	1
57	PRO.Indef.Subst.Nom.Pl.Masc	Pronome indefinido substituto, nom., pl., masc.	*beide*	1

No que concerne à frequência da combinação [caso gramatical+pessoa], os pronomes pessoais na 3.ª pessoa do nominativo ocorreram 192 vezes, revelando uma elevada utilização por parte dos informantes nas narrativas em AP em comparação com outros casos gramaticais. Por um lado, este resultado poderá sugerir uma tendência na utilização de estruturas mais simples e claras, em contrapartida, poderá indicar, também, um menor nível de proficiência linguística por parte dos informantes na utilização correta deste tipo de estruturas mais complexas e que exigem outro tipo de conhecimento. A dependência excessiva de pronomes na 3.ª pessoa pode sugerir, ainda, um vocabulário mais limitado e consequentemente pouca variação, levando a uma escrita

monótona e repetitiva. Não obstante, importa referir que muitas destas narrativas foram produzidas por crianças e, por conseguinte, a prevalência de pronomes na 3.ª pessoa pode refletir padrões específicos de desenvolvimento na aquisição da linguagem, indicando uma fase do processo de aprendizagem e não uma característica permanente do seu conhecimento linguístico.

5.6.4. Adjetivos

Com uma diferença considerável, observa-se uma forte utilização de adjetivos superlativos (*superlative adjectives*) por parte dos informantes nas suas produções em AP. A tabela 30 mostra as categorias de etiquetação identificadas. Não obstante, e ainda que em menor frequência, foram identificados adjetivos declinados de acordo com o caso, número e género.

Tabela 30

Etiquetação morfossintática da classe de palavra "Adjetivos" no subcorpus em AP.

Posição	Etiqueta	Designação	Exemplo	Frequência
1	ADJD.Pos	Adjetivo superlativo	*traurig/kaputt/böse*	200
2	ADJA.Pos.Nom.Sg.Masc	Adjetivo superlativo, nom., sing., masc.	*anderer*	12
3	ADJA.Pos.Nom.Sg.Neut	Adjetivo superlativo, nom., sing., neutro	*fröhliches*	7
4	ADJA.Pos.Acc.Sg.Neut	Adjetivo superlativo, acus., sing., neutro	*freches*	7
5	ADJA.Pos.Acc.Pl.Fem	Adjetivo superlativo, acus., pl., fem.	*bemerkte*	5
6	ADJA.Pos.Acc.Sg.Fem	Adjetivo superlativo, acus., sing., fem.	*nette*	5
7	ADJA.Pos.Nom.Sg.Fem	Adjetivo superlativo, nom., sing., fem.	*kluge*	5
8	ADJA.Pos.Dat.Sg.Masc	Adjetivo superlativo, dativo, sing., masc.	*abgefalen**	3
9	ADJA.Pos.Acc.Pl.Neut	Adjetivo superlativo, acus., pl., neutro	*hilfe**	2
10	ADJA.Comp.Nom.Sg.Fem	Adjetivo comparativo, nom., sing., fem.	*weitere*	2
11	ADJA.Pos.Acc.Pl.Masc	Adjetivo superlativo, acus., pl., masc.	*gute*	2
12	ADJA.Pos.Dat.Sg.Fem	Adjetivo superlativo, dativo, sing., fem.	*einen*	1
13	ADJA.Pos.Dat.Pl.Masc	Adjetivo superlativo, dativo, pl., masc.	*gestreckten*	1
14	ADJA.Sup.Nom.Sg.Masc	Adjetivo positivo, nom., sing., masc.	*bester*	1
15	ADJA.Pos.Dat.Sg.Neut	Adjetivo superlativo, dativo, sing., neutro	*neuen*	1
16	ADJA.Pos.Acc.Sg.Masc	Adjetivo superlativo, acus., sing., masc.	*tauchenden**	1
17	ADJA.Comp.Nom.Sg.Neut	Adjetivo comparativo, nom., sing., neutro	*grösseres*	1
18	ADJA.Sup.Nom.Pl.Masc	Adjetivo positivo, nom., pl., masc.	*beste*	1

Os resultados sugerem uma preferência pela utilização de adjetivos superlativos, dado que estes seguem, frequentemente, um padrão mais fixo em comparação com os adjetivos declinados, que requerem concordância com o género, número e caso, na língua alemã (Stocker, 2012). Ainda assim, os adjetivos superlativos no nominativo e género masculino foram os que mais ocorreram, com uma frequência de 12. Esta utilização menos frequente de adjetivos declinados, pode advir

de vários fatores que não serão devidamente aprofundados. Apenas se pretende comentar os resultados obtidos. Posto isto, a pouca utilização de formas adjetivais declinadas pode indicar, por um lado, falta de proficiência, e assim, evitar erros de concordância ou um uso estratégico por parte dos informantes em optar intencionalmente por adjetivos superlativos (não declinados) para transmitir ênfase ou exagero nas suas produções escritas. A pouca utilização de adjetivos declinados pode, também, advir do facto de os informantes se encontrarem numa fase específica do seu desenvolvimento de aquisição linguística, e não dominarem características gramaticais mais complexas, como a declinação correta de adjetivos.

5.6.5. Advérbios

A categoria dos advérbios foi uma das que teve menos frequência no subcorpus em AP. A tabela 31 mostra os resultados obtidos e onde constam alguns exemplos dos advérbios utilizados pelos informantes.

Tabela 31

Etiquetação morfossintática da classe de palavra "Advérbios" no subcorpus em AP.

Posição	Etiqueta	Designação	Exemplos	Frequência
1	ADV	Advérbio	*an*, *da*, *zu*, *nach*, *auch*, *einfach*, *aber*, *sofort*, *vorbei*, *wieder*, *dort*	235

Apesar de que estes possam ser classificados de diferentes maneiras, por exemplo, quanto ao tempo (*gestern*, *um 7 Uhr*, *manchmal*), lugar (*in die Stadt*, *dort*, *überall*), grau (*sehr*, *ziemlich*, *äußerst*), entre outras (Stocker, 2012, p. 73), essa distinção não foi feita pelo etiquetador morfossintático aquando do processamento, resultando na atribuição de apenas uma etiqueta ('ADV') para todos os advérbios presentes no subcorpus em AP.

5.6.6. Conjunções

Ao contrário da etiquetação no subcorpus em PE, observa-se, através da tabela 32, mais variedade relativamente às conjunções presentes no subcorpus em AP. No entanto, é percetível a preferência dos informantes pelo uso de conjunções coordenativas nas suas narrativas.

Tabela 32

Etiquetação morfossintática da classe de palavra "Conjunções" no subcorpus em AP.

Posição	Etiqueta	Designação	Exemplo	Frequência
1	CONJ.Coord	Conjunção coordenativa	*und*, *doch*, *denn*	155
2	CONJ.SubFin	Conjunção subordinativa finita	*weil*, *dass*, *ob*	61
3	CONJ.Comp	Conjunção comparativa	*wie*	9
4	CONJ.SubInf	Conjunção subordinativa com verbo no infinitivo	*um*	6

Como já fora referido também noutras categorias, esta preferência por conjunções coordenativas sugere uma presença de estruturas frásicas mais simples, revelando, talvez, alguma dificuldade nos informantes em dominar outras competências de sintaxe mais complexas. Não obstante, observa-se uma frequência considerável de conjunções subordinativas finitas (posição 2 da tabela 32).

5.6.7. Preposições

As preposições constituem a classe de palavras com menos frequência no subcorpus em AP. Na tabela 33 encontram-se os resultados obtidos através do processamento automático de etiquetação morfossintática.

Tabela 33

Etiquetação morfossintática da classe de palavra "Preposições" no subcorpus em AP.

Posição	Etiqueta	Designação	Exemplo	Frequência
1	APPR	Preposição	*in*, *auf*, *aus*, an	170
2	APPRART.Acc.Sg.Neut	Preposição+artigo, acus., sing., neutro	*ins*	19
3	APPRART.Dat.Sg.Masc	Preposição+artigo, dativo, sing., masc.	*zum*	7
4	APPRART.Dat.Sg.Neut	Preposição+artigo, dativo, sing., neutro	*vom*	4
5	APPRART.Dat.Sg.Fem	Preposição+artigo, dativo, sing., fem.	*zur*	2
6	APPRART.Dat.Sg.*	Preposição+artigo, dativo, sing.	*Zum*	1

Como se pode observar, a utilização de preposições, como *in*, *auf*, *aus*, *an*, entre outras, obteve uma frequência de 170. Não obstante, identificou-se a utilização de preposições em articulação com artigos, dando origem a outras formas, como é o caso do exemplo da posição 2, cuja frequência foi de 19 e denota a utilização da contração da preposição com o artigo definido no acusativo do género neutro no singular (*in* + *das* $\Rightarrow$ *ins*). Também nos exemplos das posições 3,

4, 5 e 6, foram identificadas formas de contrações de preposições com artigos (*zum*, *vom*, *zur*), embora com uma frequência ainda menor.

Em suma, a realização deste exercício de processamento do corpus e análise dos resultados obtidos ao nível morfossintático, forneceu informações sobre as características linguísticas das narrativas e sobre o desempenho dos próprios informantes. Embora a exploração dos resultados não tenha feito jus à sua complexidade e não tenha sido tão aprofundada como desejado, devido a limitações de tempo e de recursos, permitiu, no entanto, esclarecer padrões e tendências importantes sobre a proficiência linguística destas crianças. Para além disso, e apesar da presença de erros no corpus, como foi advertido ao longo da dissertação, os resultados do processamento automático de etiquetação morfossintática revelaram-se bastante elucidativos nas várias classes de palavras analisadas, oferecendo uma visão pormenorizada do comportamento linguístico e domínio dos informantes no PE e no AP.

Na secção 5.7, serão explorados e analisados os fatores sociolinguísticos que influenciam a aquisição e o uso da língua por parte destas crianças. A intersecção do desenvolvimento linguístico com os diferentes contextos de aquisição, pretende-se mostrar como se processa a aprendizagem da língua neste grupo de informantes. Embora esta dissertação represente uma exploração preliminar dos resultados, sublinha-se o potencial para futuras investigações sobre a complexidades da produção linguística dos aprendentes e a sua relação com as variáveis sociolinguísticas.

5.7. CRUZAMENTO DE VARIÁVEIS DO PERFIL SOCIOLINGUÍSTICO DOS INFORMANTES

Como já fora mencionado, os dados sociolinguísticos constituem uma componente essencial na LC, pois conferem informações relevantes sobre o percurso de aquisição, as características individuais e o contexto social e cultural no estudo linguístico. Desta forma, consegue-se compreender como a língua é moldada por fatores relacionados com a experiência linguística dos indivíduos, refletindo a interação entre o desenvolvimento linguístico e variáveis individuais e sociais como a idade, o tipo e a quantidade de contacto com as línguas, o estatuto socioeconómico, a localização geográfica, entre outras (Hansen, 2018, p. 10). Assim, a incorporação deste tipo de dados numa investigação de corpus permite uma análise profunda,

revelando padrões, variações e evoluções no uso da língua em diversos grupos sociais. De igual forma, permite-se, ainda, uma melhor compreensão da aquisição e desenvolvimento de uma ou mais línguas, especialmente em contexto migratório.

Com efeito, e tendo em consideração alguns dos fatores sociolinguísticos acima mencionados, é importante considerar a forma como estes se cruzam com iniciativas como a TEI. Como já fora também referido, a TEI oferece um quadro estandardizado para codificar e analisar dados textuais, proporcionando aos investigadores uma abordagem sistemática para documentar e explorar características linguísticas num determinado corpus. Desta forma, os fatores sociolinguísticos que acompanham as narrativas produzidas pelas crianças, encontram-se codificados de acordo com as diretrizes implementadas pela TEI.

Por sua vez, os questionários sociolinguísticos distribuídos pelos familiares dos informantes possibilitaram a recolha de inúmeros dados. Consideram-se, assim, esses dados para efeito de investigação variáveis externas (ver anexo). Na secção seguinte proceder-se-á ao cruzamento de variáveis tendo como base os dados sociolinguísticos de dois dos vinte informantes, que fazem parte da amostra de crianças em análise. Foram escolhidos apenas dois perfis de informantes para o cruzamento de variáveis, pois a realização deste exercício tornar-se-ia morosa e complexa caso fosse realizada a todo o grupo. Tendo isto em consideração, e para os efeitos de investigação nesta dissertação, a análise dos dois perfis será suficientes para a abordagem do tema e para a análise dos resultados. Desta forma, partindo das reflexões extraídas dos dois questionários, será possível estabelecer ligação com as variáveis e generalizar as observações concluídas para os restantes informantes.

5.7.1. Informante A

5.7.1.1. Contexto familiar

Os pais do informante A nasceram em Portugal, tendo emigrado para a Suíça, onde residem há 10 anos. O pai é licenciado e a mãe tem o 12.º ano concluído. O pai, responsável por ter preenchido o questionário, considera ter um excelente domínio do PE tanto na compreensão, como na produção da língua. Por oposição, o domínio do AP e da variante Suíça é apontado como "menos bom", o que significa que consegue apenas manter/perceber uma conversação simples e em ritmo lento. Em situações do quotidiano utiliza maioritariamente o PE para comunicar, à

exceção do trabalho, que refere usar tanto o PE como o AP de igual modo e quando ouve música na rádio, é exclusivamente em alemão.

Por último, considera importante que o seu filho tenha contacto com o PE, por ser vantajoso para a comunicação diária em contexto familiar e com a família alargada, assim como para o sucesso escolar do mesmo.

5.7.1.2. Perfil do informante A

Idade: 12 anos e 11 meses

Sexo: feminino

O informante nasceu em Portugal, tendo emigrado durante a infância para a Suíça aos 2 anos de idade. O contacto com o PE deu-se desde o seu nascimento através dos pais, e só mais tarde, a partir dos 3 anos de idade, é que teve o primeiro contacto com o AP no pré-escolar. O informante tem "muita fluência/quase nenhumas dificuldades" na compreensão e produção do PE e "bastante fluência/com poucas dificuldades" na compreensão e produção do AP. A comunicação com os pais é feita exclusivamente em PE. Em atividades quotidianas, como ver televisão, ouvir música e falar ao telemóvel com familiares/amigos de Portugal, o informante realiza-as quase sempre em português, sendo que a leitura de livros/revistas é, normalmente, feita em alemão.

O informante frequenta, ainda, as aulas extracurriculares de PE Língua de Herança há 6 anos, 2 horas por semana, embora não participe em atividades relacionadas com a língua e cultura portuguesas existentes na sua área de residência. Quando vai de férias a Portugal, duas vezes ao ano, comunica sempre em PE.

5.7.1.3. Resultados do processamento linguístico da narrativa

As narrativas produzidas pelo informante A foram compiladas no *Sketch Engine*, tendo sido lematizadas e etiquetadas ao nível morfossintático. O processamento destas possibilitou a recolha dos dados presentes na tabela 34.

Tabela 34

Resultados do processamento linguístico das narrativas em PE e AP do informante A.

	Narrativa em PE		Narrativa em AP	
Tokens / *Types*	123 / 63		131 / 80	
Lemas	58		63	
POS-Tagging	Categorias	Frequência	Categorias	Frequência
Verbos	9	31	9	27
Nomes	5	22	11	26
Preposições	3	8	2	7
Advérbios	2	7	1	7
Pronomes	4	5	13	18
Adjetivos	2	6	2	7
Conjunções	1	8	1	4
Numerais	-		-	

Na parte superior da tabela 34 constam os *tokens* e *types* resultantes do processamento automático de tokenização das duas narrativas, assim como os lemas obtidos através do processo de lematização, também este automático. Nota-se uma ligeira diferença entre o número de *tokens* em ambas as narrativas (123 na narrativa em PE e 131 na narrativa em AP). Observa-se, portanto, que a narrativa em AP é mais extensa que a narrativa em PE e, por conseguinte, contém uma maior quantidade de itens únicos, pois foram identificados 80 *types* na narrativa em AP e 63 *types* na narrativa em PE. Em contraste, o número de lemas identificados é bastante similar em ambas as produções escritas: 58 na narrativa em PE e 63 na narrativa em AP.

Na parte inferior da tabela 34, encontram-se expressos os resultados obtidos através do processamento automático de etiquetação morfossintática nas várias classes de palavras de ambas as narrativas. Na primeira coluna observam-se as diferentes classes de palavras (*verbos*, *nomes*, *preposições*, *etc.*), seguidas dos resultados obtidos no processo de anotação. Como é possível observar, foram identificadas 9 categorias verbais diferentes em ambas as narrativas, sendo que o número total de ocorrências variou: 31 na narrativa em PE e 27 na narrativa em AP. Tanto na narrativa em PE como na narrativa em AP, a categoria verbal mais utilizada pelo informante A foi o pretérito perfeito do indicativo na 3.ª pessoa do singular. Já na classe dos nomes, embora o número de ocorrências tenha sido, também, bastante próximo (22 frequências na narrativa em PE e 26 frequências na narrativa em AP), foram identificadas 11 categorias de nomes na narrativa em AP e 5 na narrativa em PE. Os nomes próprios no masculino singular

foram a categoria de nomes mais utilizada na narrativa em PE, enquanto os nomes próprios no nominativo singular (sem género identificado) foram a categoria de nomes mais usada na narrativa em AP.

As classes de palavras dos pronomes e das conjunções foram, também, classes onde se identificaram diferenças significativas quanto aos resultados. Nos pronomes foram identificadas 4 categorias de pronomes diferentes com uma frequência de 5 na narrativa em PE, sendo que na narrativa em AP, foram identificadas 13 categorias de pronomes e uma frequência de 18. O pronome pessoal na 3.ª pessoa do feminino no singular foi o pronome mais usado na narrativa em PE e o pronome pessoal na 3.ª pessoa do acusativo, número singular e género neutro foi o pronome mais usado pelo informante A na narrativa em AP. Por seu turno, as conjunções coordenativas foram a única categoria de conjunções usada na produção de ambas as narrativas, variando apenas na sua frequência: 8 na narrativa em PE e 4 na narrativa em AP.

Nas restantes classes de palavras os resultados foram bastante semelhantes, tendo em conta o número de categorias identificadas em cada classe de palavras e o número de ocorrências (frequência).

De salientar, ainda, que após uma análise mais detalhada por cada classe de palavra, foram identificados alguns *tokens* com erros ortográficos nas classes de palavras dos adjetivos, nomes e verbos em as ambas narrativas.

5.7.2. Informante B

5.7.2.1. Origem familiar

Os pais do informante B também nasceram em Portugal, tendo emigrado para a Suíça, onde residem há 22 anos. Ambos têm apenas o 2.º ciclo de escolaridade. A mãe, responsável pelo preenchimento do questionário, considera ter um excelente domínio do PE tanto na compreensão, como na produção da língua e um bom domínio do AP e da variante Suíça, uma vez que consegue manter/perceber conversações mais longas. A utilização de ambas as línguas está muito presente no seu quotidiano, com exceção no contacto com familiares e amigos portugueses residentes na Suíça, assim como em clubes/associações portuguesas e para ver televisão, que é apenas em PE. Considera, ainda, importante que o seu filho tenha contacto com o PE.

5.7.2.2. Perfil do informante B

Idade: 11 anos e 7 meses

Sexo: masculino

O informante nasceu na Suíça. O contacto com o PE deu-se desde o seu nascimento através dos pais e irmão mais velho (que também nasceu na Suíça), e só mais tarde, a partir dos 6 anos de idade, é que teve o primeiro contacto com o AP na escola primária. O informante tem "fluência nativa/sem nenhuma dificuldade" na compreensão e produção do PE e do AP. A comunicação com os pais é feita exclusivamente em PE e com o irmão mais velho é feita metade em português e metade em alemão (variante suíça). Em atividades quotidianas, como ouvir música e ler livros/revistas, o informante realiza-as normalmente em AP, sendo que falar ao telemóvel com familiares/amigos de Portugal é normalmente metade em português e metade em alemão (variante suíça).

O informante frequenta, ainda, aulas extracurriculares de PE Língua de Herança há 5 anos, 1H30 por semana, embora não participe em atividades relacionadas com a língua e cultura portuguesas existentes na sua área de residência. Quando vai de férias a Portugal, duas vezes ao ano, comunica sempre em PE.

5.7.2.3. Resultados

De acordo com os resultados expressos na tabela 35, as narrativas apresentam uma diferença considerável quanto ao número de *tokens*, mostrando que a narrativa em PE é mais extensa (215 *tokens*) que a narrativa em AP (137 *tokens*). Por conseguinte, também o número de lemas, bem como o número de categorias e frequência em algumas classes de palavras, apresentam diferenças consideráveis entre as duas narrativas.

Tabela 35

Resultado do processamento linguístico das narrativas em PE e AP do informante B.

	Narrativa em PE		Narrativa em AP	
Tokens / Types	215 / 109		137 / 80	
Lemas	89		65	
POS-Tagging	Categorias	Frequência	Categorias	Frequência
Verbos	11	48	8	25
Nomes	5	38	13	24
Preposições	3	29	2	10
Advérbios	2	11	1	12
Pronomes	8	18	11	15
Adjetivos	1	6	3	8
Conjunções	2	16	4	10
Numerais	-		-	

A classe dos verbos é a classe que mais discrepância apresenta relativamente ao número de ocorrências quando comparadas as narrativas em PE e AP. Na narrativa em PE foram identificadas 11 conjugações verbais diferentes e registada uma frequência de 48, sendo que na narrativa em AP, o número de conjugações verbais diferentes foi ligeiramente mais baixo, nomeadamente 8, enquanto a frequência foi de 25. Por conseguinte, e sem grande margem para dúvida, tanto na narrativa em PE como na narrativa em AP, a conjugação verbal mais utilizada pelo informante B foi o pretérito perfeito do indicativo na 3.ª pessoa do singular. Também na classe dos nomes se notou uma diferença considerável, tendo sido identificada uma frequência de 38 nomes na narrativa em PE e uma frequência de 24 na narrativa em AP. Relativamente às categorias, na narrativa em PE identificaram-se 5 categorias de nomes diferentes e 13 na narrativa em AP. Já a frequência da classe dos nomes em PE foi de 38 e de 24 na narrativa em AP. Outra classe de palavras onde também se notou uma grande discrepância ao nível do número de ocorrências foi na classe das preposições. Na narrativa em PE foram identificadas 29 preposições e 10 na narrativa em AP. Quanto às diferentes categorias de preposições, a diferença foi relativamente baixa, tendo sido identificadas 3 categorias na narrativa em PE e 2 na narrativa em AP.

Os resultados obtidos nas restantes classes de palavras mantiveram-se, de certa forma, semelhantes a nível das proporções entre categorias e ocorrências em ambas as narrativas.

Relativamente à presença de erros ortográficos, foram identificados apenas erros pontuais na classe de palavras dos nomes e pronomes na narrativa em PE e na classe dos verbos na narrativa em AP. Por último, de referir que não foram identificados erros na classe de palavras dos verbos na narrativa em PE.

5.7.3. Discussão de resultados

Numa primeira instância, os informantes A e B apresentam características muito semelhantes quanto ao seu processo de aquisição linguística. Primeiramente, em ambos os casos, o contacto com a língua e cultura do país de acolhimento ocorreu numa fase precoce, tendo o informante A emigrando durante a infância com apenas 2 anos, e o informante B nascido na Suíça. Apesar desta exposição precoce à língua e cultura do país de acolhimento, em ambas situações, o PE foi a primeira língua à qual foram expostos, tendo sido mais tarde, no pré-escolar e na primária, respetivamente, que tiveram o primeiro contacto mais intensivo com o AP. De igual forma, ambos encarregados de educação referiram utilizar a língua de herança na comunicação diária, com exceção do informante B, que por vezes também comunica em alemão suíço com o seu irmão mais velho. Não obstante, apesar dos informantes não participarem em atividades culturais em associações portuguesas na sua área de residência, o contacto com a língua e cultura de herança, para além de ser mantido no seio familiar e com amigos e familiares portugueses à distância, ambos informantes frequentam aulas extracurriculares de PE e deslocam-se duas vezes por ano a Portugal. Para além disso, também em vários momentos de lazer, como ver televisão, ler ou ouvir música, é mantido por estas crianças um contacto com o PE. Dada esta conjuntura de fatores, percebe-se que, embora tenham emigrado há vários anos, o contacto com Portugal e a língua e cultura portuguesas nunca deixou de existir, fazendo, inclusive, parte do quotidiano destas crianças. Ademais, esta exposição diária a que estão sujeitos os informantes contribui bastante para que tivessem adquirido e consigam manter um determinado nível de proficiência equiparado ao de um nativo, como aliás, se percebe através das suas produções escritas.

No que concerne às narrativas, é, também, percetível através dos dados recolhidos, que ambos informantes conseguiram demonstrar um nível de proficiência linguística muito idêntico em ambas as línguas, fruto de todos os fatores que contribuem para que a exposição à língua de herança permaneça e seja constante. Também Flores, Rinke *et al.* (2022), através do seu estudo sobre a produção da morfologia verbal e da posição dos verbos num corpus produzido por 60 informantes

bilingues (português-alemão), concluíram que os informantes apresentam conhecimentos sintáticos e morfológicos estáveis em ambas as línguas. Os dados recolhidos para o estudo de Flores, Rinke *et al*. (2022) são os mesmos utilizados no âmbito desta dissertação. Os autores apontam, ainda, que apenas foram identificadas diferenças de desempenho entre o corpus português e o alemão ao nível da ortografia. Relativamente às narrativas produzidas pelos informantes A e B, existe um melhor desempenho na narrativa em AP, como é de esperar, tendo em conta que o AP é a língua de escolarização dos informantes e tem um peso significativo no país de acolhimento, nomeadamente nas instituições de ensino, apesar de no registo oral, a variante suíça ser a mais utilizada. Ainda assim, a diferença que se fez sentir através dos resultados é ligeira. Por outro lado, apesar de terem sido identificados alguns erros ortográficos, estes também não têm um peso significativo na análise global do desempenho dos informantes nas suas produções escritas, demonstrando, uma vez mais, que os níveis de proficiência linguística entre o PE e o AP são bastante equiparáveis. Neste sentido, também os resultados obtidos nesta dissertação vão de encontro aos resultados do estudo de Flores, Rinke *et al*. (2022).

CAPÍTULO VI – PROTÓTIPO DE INTERFACE DE PESQUISA

Este capítulo representa a segunda parte desta dissertação, uma vez que se destina à apresentação de uma prova-conceito de um protótipo de interface de pesquisa e constituirá um exercício de reflexão sobre o papel fundamental que este tipo de recurso tem nas investigações linguísticas. Com a crescente disponibilidade e importância de (grandes) corpora em vários domínios da linguística, o papel das ferramentas, como *softwares* e interfaces de pesquisa, passou a ser considerado como um instrumento científico parte integrante dos processos de investigação (Geyken & Kupietz, 2016). Não obstante, devido a limitações de tempo e recursos, não se procederá à elaboração de uma interface de pesquisa, mas sim à exposição de alguns aspetos, tais como conceitos, objetivos, público-alvo, funcionalidades, entre outros elementos relevantes para uma futura elaboração e implementação de um protótipo de interface, que colmate as necessidades deste e outros projetos de investigação.

6.1. OBJETIVOS DA PROVA-CONCEITO

Como fora já mencionado, o recurso a amostras de linguagem, através da criação e compilação de corpora, tem crescido ao longo dos anos. Nesse sentido, pretende-se que este e outros tipos de recursos semelhantes não fiquem restringidos aos projetos para os quais foram criados, contribuindo para a sua sustentabilidade e eficácia no âmbito da linguística. Soehn *et al.* (2008) referem que "the process of building a language resource is expensive, time-consuming, and it includes aspects such as corpus sampling and linguistic annotation on multiple levels" (p. 27). Desta forma, a questão da sustentabilidade e da preservação de corpora tem vindo a ganhar visibilidade dentro da comunidade científica, mostrando, cada vez mais, um certo cuidado com a otimização destes recursos, principalmente findados os projetos para os quais foram recolhidos e/ou criados. Como Soehn *et al.* (2008) afirmam, dada a morosidade, a mobilização de recursos e os custos associados à recolha de dados linguísticos, em particular aos dados utilizados neste projeto de dissertação, recolhidos por investigadoras portuguesas a informantes que vivem na Suíça, torna-se pertinente a questão da sustentabilidade, preservação e do reaproveitamento deste tipo de recursos para outros fins investigativos. Desta forma, a criação deste protótipo de interface de pesquisa tem como objetivo principal o armazenamento do corpus criado para este projeto de dissertação. Assim, pretende-se que, numa primeira instância, esta interface funcione como um

repositório e/ou base de dados, que possa ser disponibilizada e acessível à comunidade científica, professores, familiares dos informantes, entre outros utilizadores que possam ter interesse por esta área de conhecimento e investigação.

Pretende-se, assim, que o armazenamento do corpus na interface de pesquisa esteja disponível para visualização e/ou *download* nos formatos .txt e .xml, tanto na sua versão não anotada, como também na sua versão anotada. Para a visualização dos dados, almeja-se que o utilizador possa aceder ao corpus e às respetivas camadas de anotação sem a necessidade de o ter de descarregar, bem como a realização de pesquisas no próprio corpus (por exemplo, efetuar uma pesquisa através de um lema e verificar as flexões associadas ao mesmo). A preservação do corpus na sua versão não anotada, possibilitaria, numa segunda instância, que outros utilizadores, ao descarregarem os ficheiros do corpus, possam utilizar essas amostras para fins de investigação, explorando, por exemplo, outros níveis de anotação, como o ortográfico, dado que os erros produzidos pelos informantes foram preservados. Por outro lado, a versão anotada do corpus ao nível morfossintático e a sua lematização, possibilitam, também, uma panóplia de características e padrões suscetíveis para análise e futuras investigações, diferentes das que foram indagadas nesta dissertação.

Para além do objetivo da sustentabilidade e reaproveitamento dos dados linguísticos, também a metainformação recolhida no âmbito dos questionários sociolinguísticos seria disponibilizada e acessível através da interface, dado a relevância que apresenta nesta área de estudos, possibilitando o cruzamento de variáveis e apresentação de novas conclusões.

Relativamente ao tamanho do corpus, perspetiva-se que este possa ser alargado, uma vez que neste projeto de dissertação foram utilizadas apenas 40 narrativas das cerca de 500 recolhidas no âmbito do projeto *Competência bilingue de crianças lusodescendentes residentes na Suíça*. Desta forma, as restantes narrativas poderiam, também, ser etiquetadas ao nível morfossintático e lematizadas, ou ainda, processadas a outros níveis com recurso a uma ferramenta de processamento de corpora. A criação desta interface de pesquisa possibilita não só a preservação de amostras da linguagem recolhidas em contexto real, como a sua difusão, e ainda, que estes dados, ao serem reutilizados e reaproveitados, possam servir de objeto de estudo para futuros trabalhos de pesquisa.

Para a apresentação da prova-conceito do protótipo de interface não serão tidos em consideração detalhes direcionados ao aspeto visual, pois nesta primeira fase, pretende-se a apresentação de

um protótipo acima de tudo funcional, centrado nas opções de pesquisa e que cumpra com os requisitos estabelecidos. A sua estrutura manter-se-á, portanto, numa linha objetiva e minimalista quanto aos detalhes gráficos.

Para ajudar ao processo de reflexão e apresentação da prova-conceito do protótipo de interface de pesquisa, apresenta-se, no ponto 6.2., alguns projetos com exemplos de interfaces de pesquisa de relevância, onde serão mencionadas as principais características.

6.2. INTERFACES DE PESQUISA DE CORPORA EXISTENTES

De acordo com Huang e Li (2010), "a well-designed user interface entails carefully considering the particular user group of the application and delivering an application that works effectively and efficiently" (p. 352). Os autores realçam a importância de uma conceção cuidadosa na criação de uma interface de pesquisa, afirmando que se deve ter sempre em consideração o grupo de utilizadores e compreender as suas necessidades, preferências e comportamentos, de forma a proporcionar uma boa experiência e que esta corresponda às expectativas dos mesmos. Huang e Li (2010) sublinham que na criação de uma interface de pesquisa, "one needs to make sure that the interface matches the way users want to accomplish a task. One also needs to use the most appropriate modality at the appropriate time to assist users to achieve their goals" (p. 352).

Neste sentido, apresentam-se alguns projetos que levaram a cabo a criação de interfaces de pesquisa e/ou repositórios, com o objetivo de contribuir para a reutilização de recursos linguísticos de forma acessível a grupos de utilizadores específicos. Os exemplos que se encontram abaixo mencionados dispõem de características e funcionalidades que se pretendem incluir no protótipo de interface de pesquisa descrito nesta dissertação. Desta forma, destacam-se projetos que contribuem para o reaproveitamento dos vários recursos linguísticos numa base de dados, que estes estejam disponíveis através de uma interface intuitiva e que permita ao utilizador efetuar pesquisas tanto nos corpora como através dos metadados.

6.2.1. CROW

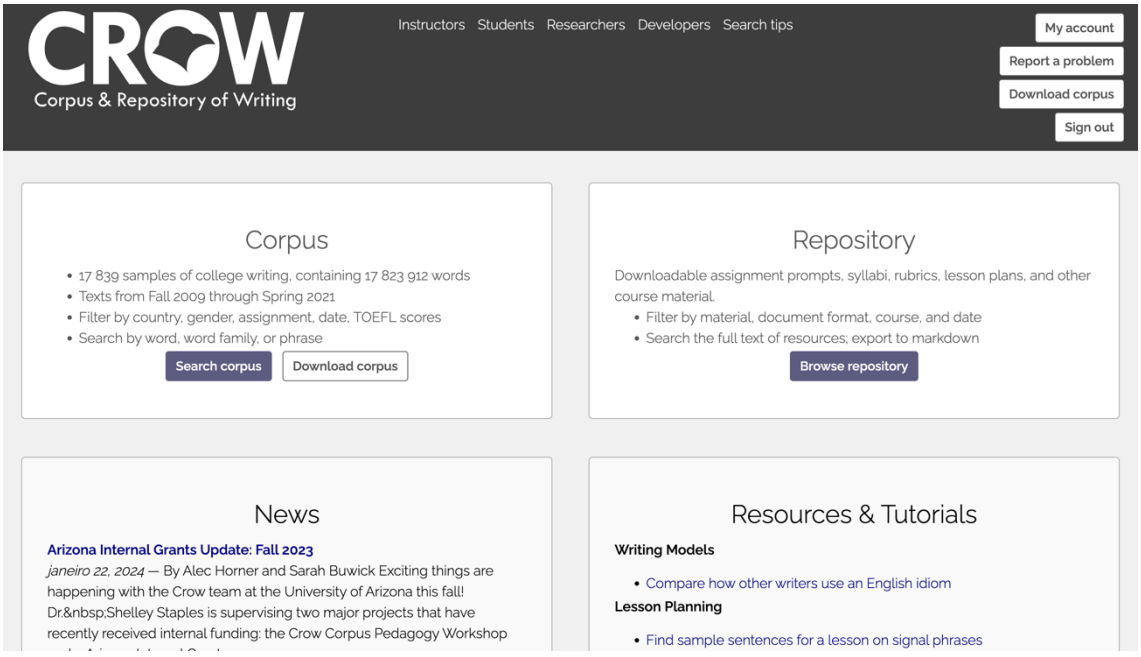
O *Corpus & Repository of Writing*[68] (CROW) consiste numa base de dados de acesso livre que alberga um corpus de aprendizagem composto por mais de 10.000 textos produzidos por

[68] https://crow.corporaproject.org/

estudantes universitários oriundos de mais de 50 países diferentes (Yaylali et al., 2021, p. 2). À semelhança do corpus criado para esta dissertação, também os dados disponíveis no CROW correspondem a dados de aprendentes de uma L2. A criação do projeto ocorreu em 2014 na Universidade de Purdue (EUA), tendo-se expandido para mais cinco instituições de ensino superior nos EUA (Lan et al., 2019, p. 106), com o objetivo de apoiar a investigação e o desenvolvimento profissional em linguística aplicada e em retórica e composição. Lan *et al.* (2019) explicam "the learner corpus in Crow consists of writing assignments written by L1 and L2 students in first-year composition courses" (p. 106). Para aceder ao CROW, o utilizador tem de se registar na plataforma e efetuar um pedido de acesso para a visualização dos dados e manuseamento da interface. A ilustração 5[69] mostra a página inicial do CROW, depois de autorizado o acesso e colocando as credenciais (*user* e palavra-passe) escolhidas pelo utilizador aquando do registo.

Ilustração 5

Página inicial da base de dados do CROW.



Como é possível observar através da ilustração 5, depois de ingressar na página inicial do CROW, o utilizador pode aceder diretamente ao corpus para efetuar pesquisas e/ou proceder ao *download*

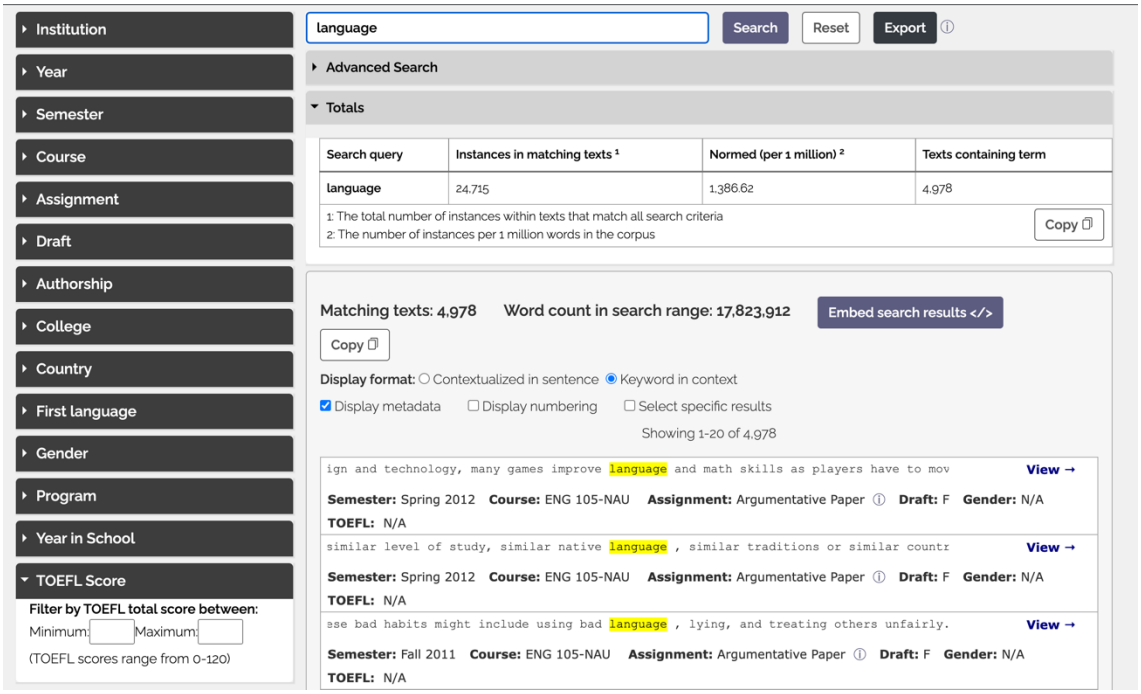
[69] O acesso à base de dados CROW realizou-se em janeiro de 2024 e, por conseguinte, as capturas dessa interface que se encontram nesta dissertação foram obtidas também nesse período.

do mesmo. Através do *download*, os dados (não anotados) são disponibilizados em formato .txt, com a devida metainformação referente ao informante e encontram-se organizados por pastas de acordo com o nível de proficiência e tipo de produção escrita. No repositório, o utilizador tem, ainda, acesso a materiais de apoio à sua investigação, como atividades, tipos de exames, rúbricas, entre outros documentos relevantes à pesquisa linguística.

Relativamente à pesquisa no corpus, a ilustração 6 mostra os resultados obtidos aquando da pesquisa da palavra "language".

Ilustração 6

Resultados obtidos na pesquisa do token "language" no corpus CROW.



Como é possível visualizar através da ilustração 6, encontra-se, na parte superior da plataforma, uma barra de pesquisa onde é possível introduzir a palavra que o utilizador pretenda pesquisar. Por sua vez, do lado esquerdo, a cinzento-escuro, aparecem vários filtros, como a instituição de ensino, o ano em que o texto foi produzido, o país de onde o informante é oriundo, qual a sua L1, entre outros elementos, que o utilizador pode selecionar se pretender efetuar uma pesquisa mais específica. Na parte central da plataforma são visíveis os resultados obtidos através da pesquisa. Na pequena tabela, logo abaixo da barra de pesquisa, aparecem alguns números referentes à

quantidade de vezes que a palavra pesquisada aparece nos textos de acordo com os critérios selecionados e também por milhão de palavras no corpus, e ainda, a quantidade de textos que contém a palavra pesquisada. Mais abaixo, é possível uma visualização dos resultados em linhas de concordância, para que o utilizador tenha acesso ao contexto em que a palavra se insere. Para além disso, o utilizador pode ter acesso à metainformação, que aparece abaixo da linha de concordância.

É possível, ainda, a realização de pesquisas mais avançadas, as quais não se encontrarão descritas nesta dissertação e ficarão suscetíveis de uma indagação ao critério do utilizador. Pretende-se apenas mostrar, de forma sucinta, as principais característica desta e das restantes interfaces, que são relevantes para esta dissertação.

6.2.2. COSMAS II

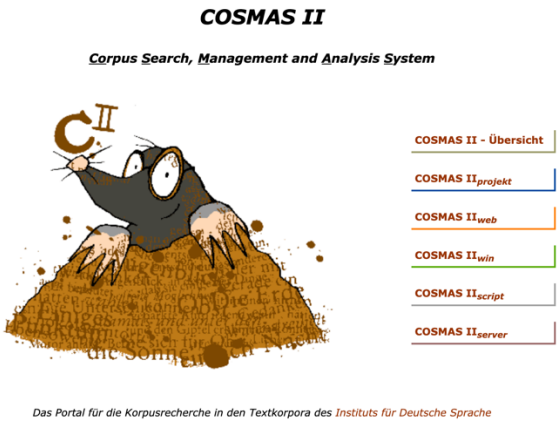
O *Corpus Search, Management and Analysis System II*[70] (COSMAS II) foi desenvolvido pelo Leibniz-Institut für Deutsche Sprache (IDS), sucedendo ao sistema anterior COSMAS I (1991-2003), e constitui uma interface de pesquisa, que possibilita a consulta de grandes quantidades de dados textuais tendo em conta a anotação baseada em palavras (Soehn et al., 2008, p. 28). Assim, através desta interface, o utilizador pode efetuar pesquisas em subcorpora por via de metainformação ou extrair dados que contenham expressões ou palavras previamente pesquisadas (Soehn et al., 2008, p. 29). Os resultados são apresentados sob a forma de linhas de concordância. O utilizador pode, também, obter informações sobre o TTR e informação estatística sobre as colocações. A ilustração 7[71] mostra a página inicia da plataforma COSMAS II.

[70] http://www2.ids-mannheim.de/cosmas2/

[71] O acesso ao sistema COSMAS II realizou-se em janeiro de 2024 e, por conseguinte, as capturas dessa interface que se encontram nesta dissertação foram obtidas também nesse período.

Ilustração 7

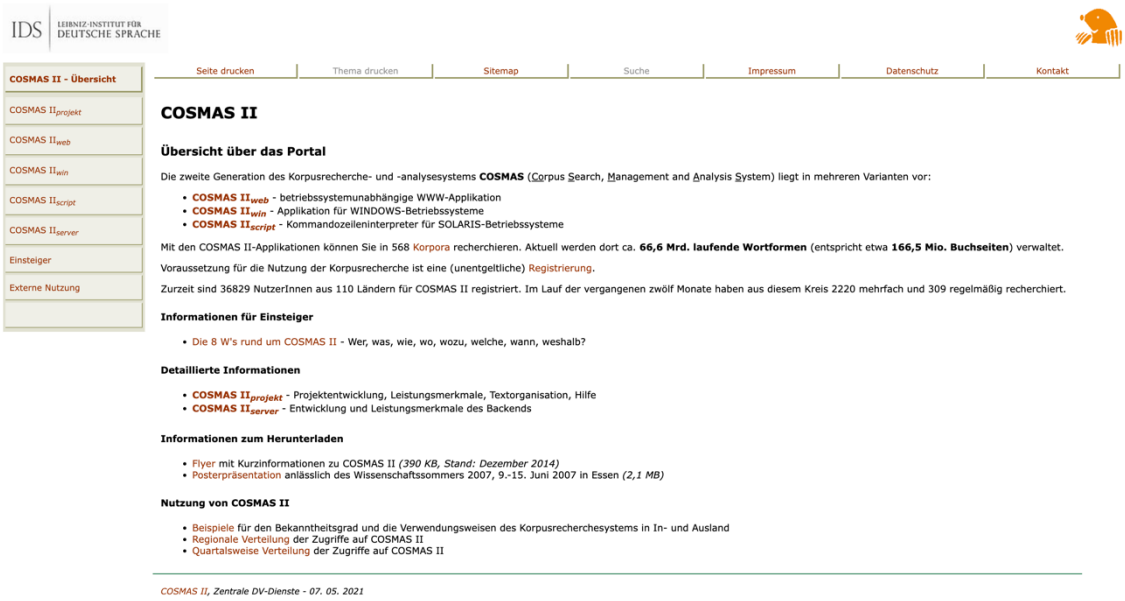
Página inicial da interface COSMAS II.



Ao ingressar na página inicial do COSMAS II (ilustração 7) é possível, através do menu disponível do lado direito, o acesso a uma visão mais pormenorizada da página ao clicar no separador "COSMAS II – Übersicht" (ilustração 8). Desta forma, é fornecido ao utilizador um panorama da interface, do projeto (desenvolvimento, características e organização) e dos sistemas de pesquisa disponíveis para a exploração e análise do corpus COSMAS II.

Ilustração 8

Separador "Übersicht" (visão geral) da interface COSMAS II.

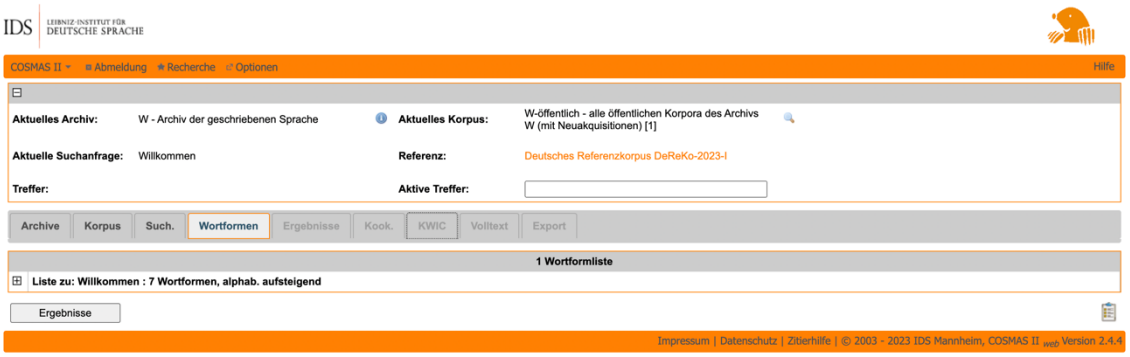


Também à semelhança do CROW, é necessário que o utilizador se registe (gratuitamente) para poder aceder aos dados disponíveis. Através da interface COSMAS II é possível consultar e manusear o corpus de referência DeReKo, assim como outros corpora históricos e não-históricos.

Como se pode observar na ilustração 8, existem dois menus de navegação colocados na parte superior e na margem lateral direita da interface, que permitem ao utilizador navegar pela interface de forma mais eficiente e intuitiva. Relativamente à pesquisa, o utilizador pode efetuar vários tipos de pesquisa no corpus, uma vez que já se encontra registado. São disponibilizados vários subcorpora onde o utilizador pode efetuar, por exemplo, pesquisas através da forma exata da palavra (*Wortformen*), como é possível visualização pela ilustração 9.

Ilustração 9

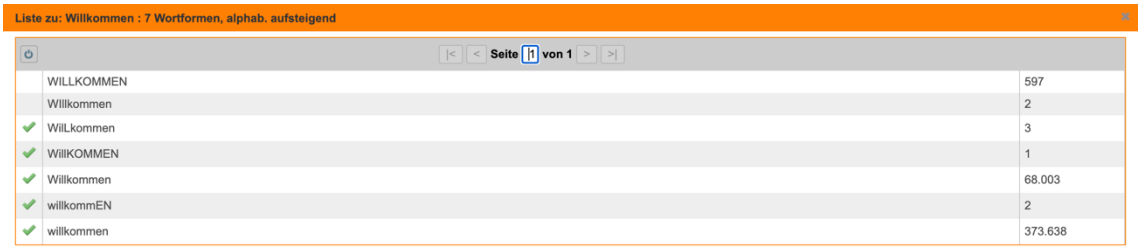
Apresentação da funcionalidade de pesquisa por palavra (Wortformen) na interface COSMAS II.



Para efetuar a pesquisa, utilizou-se a palavra "willkommen", que em português significa "bem-vindo". Os resultados da pesquisa são demonstrados através de listas de palavras, à semelhança da ferramenta *Sketch Engine*, onde se observam as várias formas ortográficas que a palavra pesquisada adquire no subcorpus e a quantidade de vezes que estas ocorrem (ilustração 10).

Ilustração 10

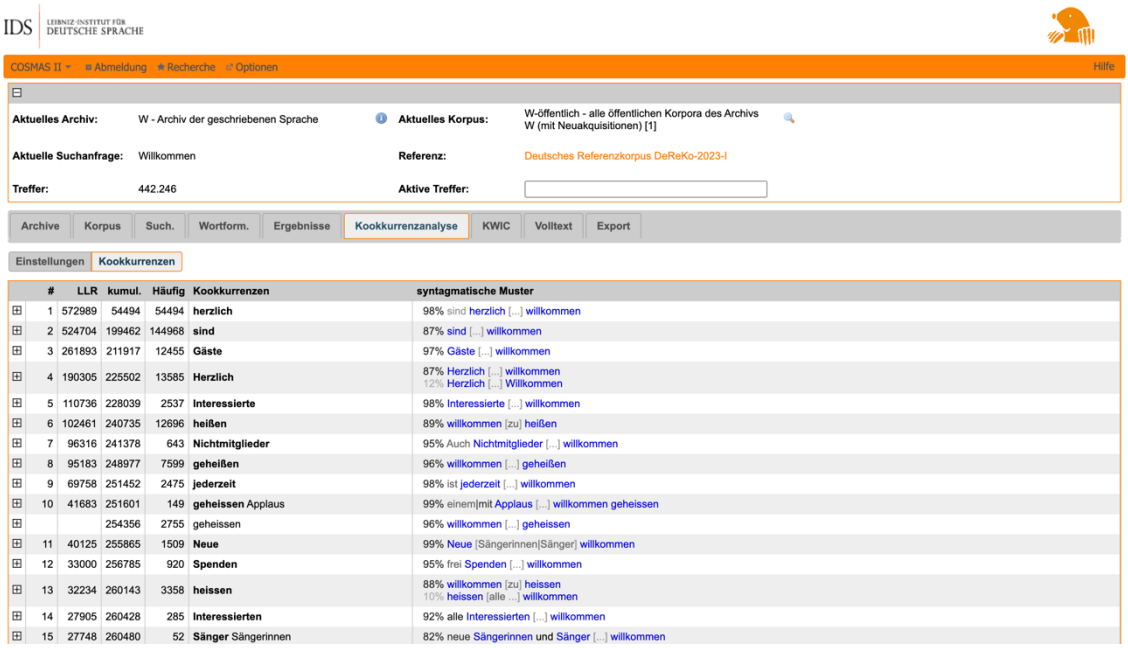
Lista de palavras obtida através da pesquisa da palavra "willkommen" onde é possível observar as várias formas ortográficas e sua a frequência.



Para além disso, o utilizador pode efetuar uma análise de co-ocorrências no corpus e verificar, assim, as estruturas associadas à palavra pesquisada. Algumas das estruturas mais utilizadas com a palavra "willkommen" são, como é possível observar através da ilustração 11, "herzlich...willkommen" com 98% de ocorrência, "sind...willkommen" com 87% de ocorrência, "Gäste...willkommen" com 97% de ocorrência, entre outras.

Ilustração 11

Resultado da co-ocorrência obtida através da pesquisa da palavra "willkommen" na interface COSMAS II.

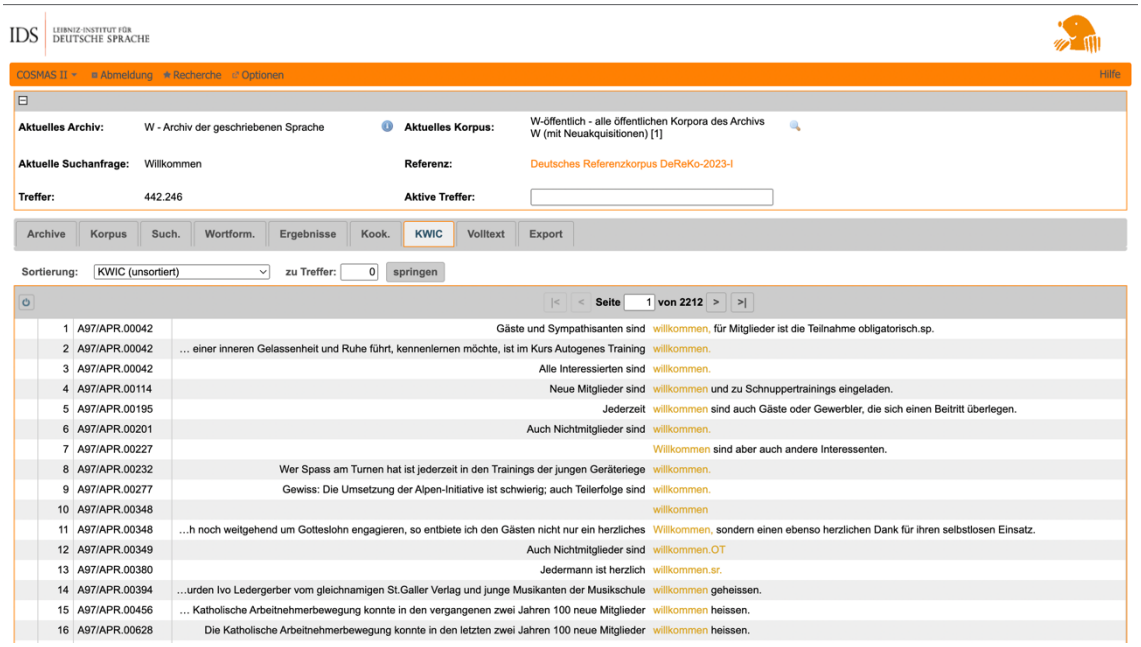


#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
1	572989	54494	54494	**herzlich**	98% sind herzlich [...] willkommen
2	524704	199462	144968	**sind**	87% sind [...] willkommen
3	261893	211917	12455	**Gäste**	97% Gäste [...] willkommen
4	190305	225502	13585	**Herzlich**	87% Herzlich [...] willkommen 12% Herzlich [...] Willkommen
5	110736	228039	2537	**Interessierte**	98% Interessierte [...] willkommen
6	102461	240735	12696	**heißen**	89% willkommen [zu] heißen
7	96316	241378	643	**Nichtmitglieder**	95% Auch Nichtmitglieder [...] willkommen
8	95183	248977	7599	**geheißen**	96% willkommen [...] geheißen
9	69758	251452	2475	**jederzeit**	98% ist jederzeit [...] willkommen
10	41683	251601	149	**geheissen** Applaus	99% einem|mit Applaus [...] willkommen geheissen
		254356	2755	geheissen	96% willkommen [...] geheissen
11	40125	255865	1509	**Neue**	99% Neue [Sängerinnen|Sänger] willkommen
12	33000	256785	920	**Spenden**	95% frei Spenden [...] willkommen
13	32234	260143	3358	**heissen**	88% willkommen [zu] heissen 10% heissen [alle ...] willkommen
14	27905	260428	285	**Interessierten**	92% alle Interessierten [...] willkommen
15	27748	260480	52	**Sänger** Sängerinnen	82% neue Sängerinnen und Sänger [...] willkommen

Por fim, o utilizador pode, ainda, efetuar a sua pesquisa no corpus de acordo com o contexto, através da funcionalidade KWIC. A ilustração 12 mostra a palavra "willkommen" de acordo com o contexto em que é utilizada no corpus.

Ilustração 12

Resultado da pesquisa da palavra "willkommen" de acordo com o contexto, através da funcionalidade KWIC.



O COSMAS II representa um sistema bastante completo e benéfico para a gestão de grandes conjuntos de dados linguísticos, facilitando o seu acesso e organização. Além disso, oferece recursos de pesquisa avançados, permitindo que o utilizador realize consultas linguísticas complexas com precisão e eficiência.

6.2.3. KonText

O KonText[72] (Machálek, 2020) é uma interface de pesquisa de corpora altamente personalizável construída com base nas bibliotecas principais do motor de pesquisa de corpus de código aberto *NoSketch Engine* (NoSkE)[73] (p. 7003). Segundo Machálek (2020) "the aim is to overcome some limitations of the original NoSkE user interface and provide integration capabilities allowing connection of the basic search service with other language resources (LRs)" (p. 7003). Encontra-

[72] https://www.korpus.cz/kontext/query?corpname=syn2020

[73] O *NoSketch Engine* é uma versão de fonte aberta do *Sketch Engine*, porém com certas limitações de funcionalidade. Em contraste com a ferramenta utilizada para o processamento do corpus em estudo, no NoSkE não existem corpora pré-carregados, não existe uma interface gráfica que permita fazer a gestão do corpus, não existem ferramentas para criação de um corpus, como a tokenização automática, lematização, POS-Tagging, não havendo, ainda, funcionalidades como o *Word Sketches*, *Thesaurus*, *N-grams*, entre outras. Desta forma, o utilizador deve possuir conhecimentos técnicos adequados e utilizar ferramentas externas para o processamento dos corpora.
Fonte: https://www.sketchengine.eu/nosketch-engine/

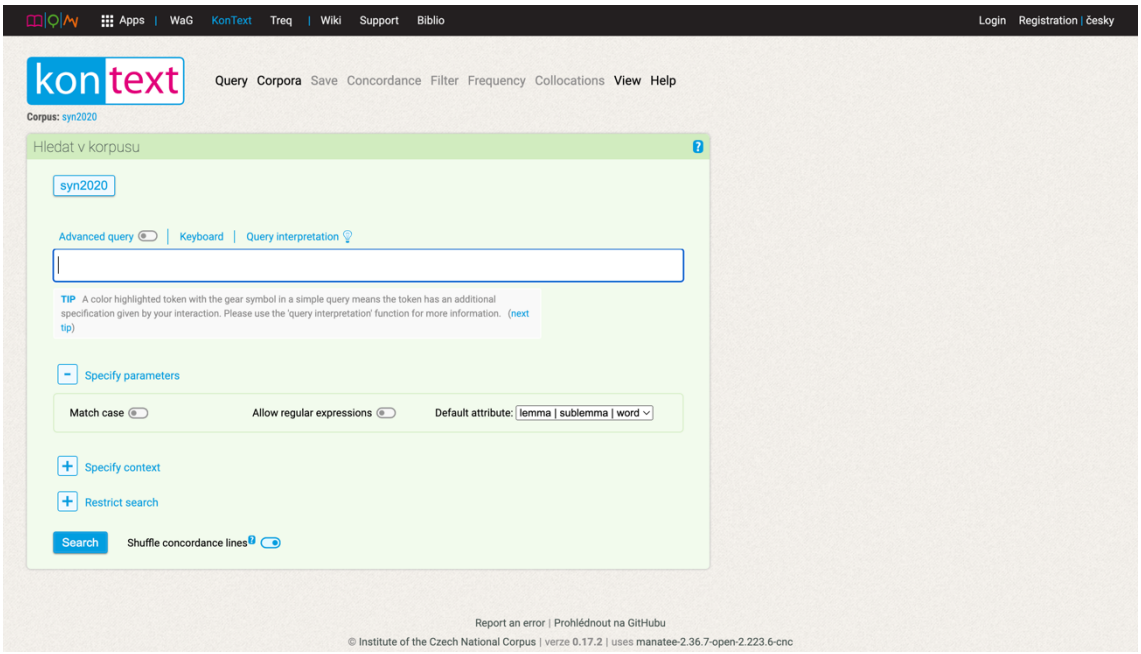
se operacional desde 2014 e contou com a colaboração de vários utilizadores e investigadores do *Czech National Corpus* (CNC) (Machálek, 2020, p. 7003). De acordo com Machálek (2020):

> KonText is endowed with the following key features: support for spoken (audio playback, visual representation of dialogues) and parallel corpora; rendering of dependency syntax trees; advanced creation of subcorpora: based on user-defined ratios of different text types and based on the corpora alignment; helper tools for query creation and exploring of corpora structure; and integration with other services that can provide additional information about searched terms. (p. 7003)

Graças a ferramentas incorporadas no KonText, é possível carregar e pesquisar os próprios corpora do utilizador, que são automaticamente anotados ao nível morfossintático. A sua estrutura e as funcionalidades que oferece assemelham-se, em parte, às funcionalidades oferecidas pelo *Sketch Engine*, onde o utilizador pode efetuar pesquisas por concordância, obter distribuição de frequências e proceder à análise de colocações. A ilustração 13[74] mostra a página inicial desta interface.

Ilustração 13

Página inicial da interface KonText.

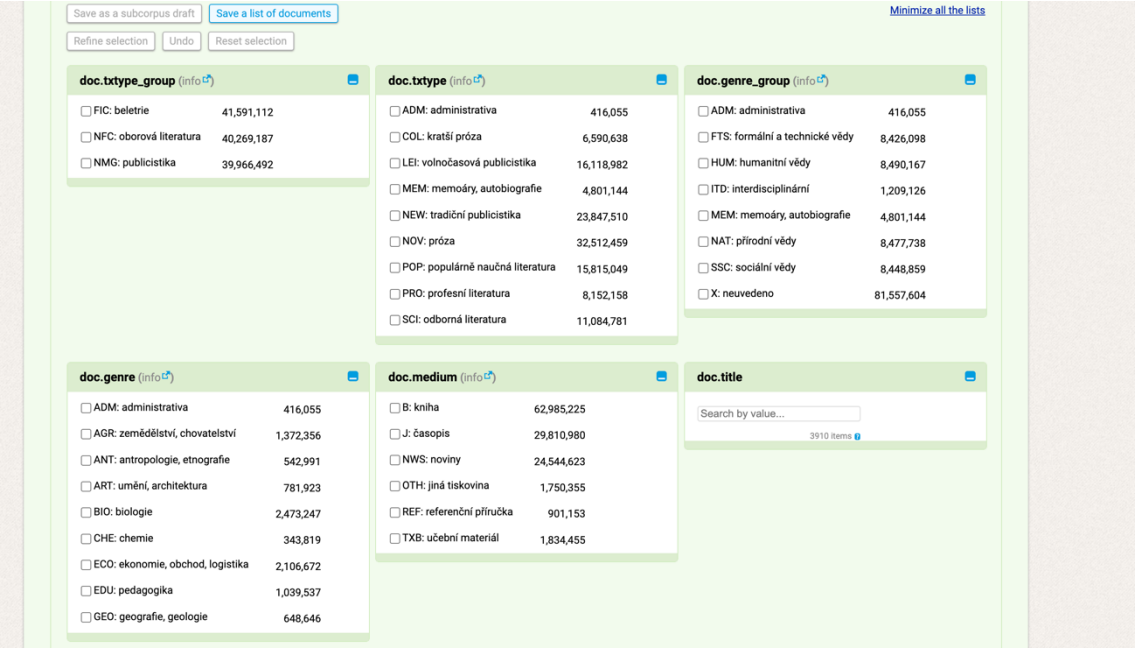


[74] O acesso à interface KonText realizou-se em janeiro de 2024 e, por conseguinte, as capturas dessa interface que se encontram nesta dissertação foram obtidas também nesse período.

A página inicial mostra uma barra de pesquisa, onde o utilizador pode introduzir a palavra que pretende pesquisar e os resultados aparecerão em formato de KWIC. Das várias plataformas que foram exploradas e apresentadas nesta dissertação, a apresentação dos resultados no formato KWIC tem sido bastante comum em interfaces que alojam corpora e permitem efetuar pesquisas linguísticas deste tipo. Nas secções abaixo da barra de pesquisa (*Specify parameters*, *Specify context* e *Restrict search*), o utilizador pode adicionar filtros, como o género do corpus, o ano de publicação, a temática, o tipo de texto, entre outras características, para personalizar a sua pesquisa. Atente-se na ilustração 14.

Ilustração 14

Filtros para personalizar a pesquisa na interface KonText.



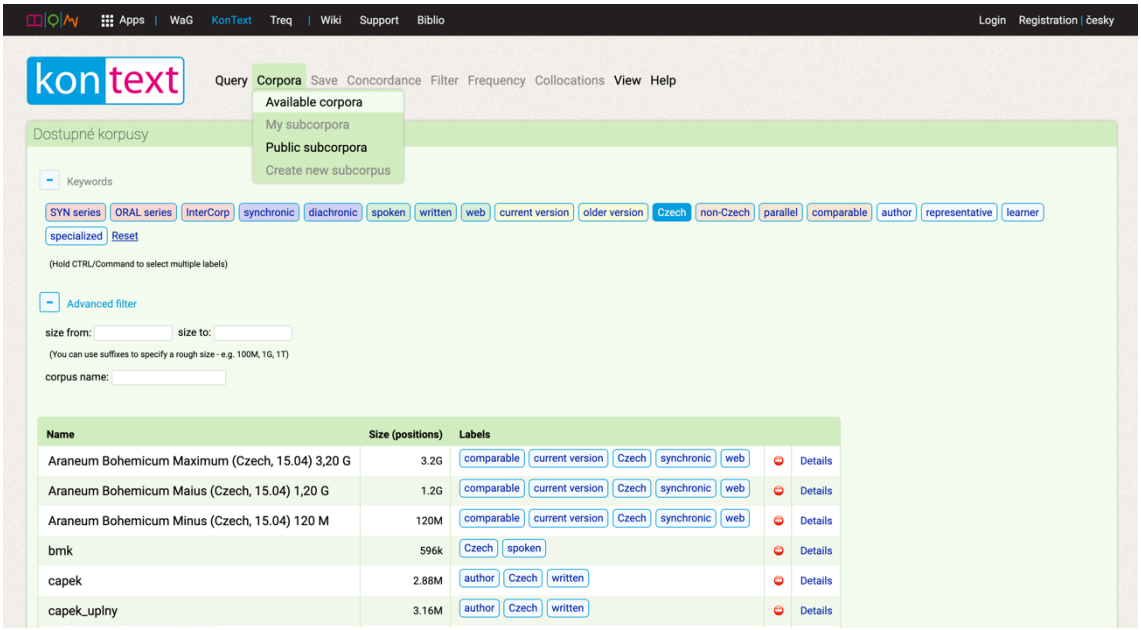
Para além disso, o facto de a interface estar integrada no projeto do CNC, permite que o utilizador possa ter acesso a um conjunto alargado de outras aplicações para a exploração de questões linguísticas. Destacam-se, por exemplo, o *Word at a Glance* (WaG), que tem como objetivo criar uma visão geral da utilização de uma palavra, a ferramenta *SyD*, que foi concebida para a exploração versátil de variantes, tanto do ponto de vista sincrónico (língua contemporânea) como diacrónico e a ferramenta *Morfio*, onde o utilizador pode procurar relações de formação de

palavras entre unidades do corpus, encontrando todos os pares de palavras formados da mesma maneira e avaliar a produtividade morfológica da sua formação.

Através da lustração 15, é possível visualizar as funcionalidades disponíveis na interface, através do seu menu superior, assim como os corpora pré-carregados, que podem servir de base para pesquisas. Desta forma, é possível aceder, entre outros aspetos, ao nome do corpus, ao tamanho e a algumas características do mesmo, através de etiquetas que o identificam.

Ilustração 15

Funcionalidades de pesquisa disponíveis na interface KonText.



Para além das características herdadas do NoSkE, o KonText oferece um conjunto de características originais, que raramente se encontram noutras interfaces de pesquisa de corpus, onde se destacam a seleção interativa de texto, seleção de texto baseada na proporção e a distribuição de frequências bidimensional (Machálek, 2020, pp. 7004-7005). Segundo o autor, um módulo de seleção interativa de texto (com base numa combinação de critérios) facilita a criação de subcorpora personalizados pelo utilizador, possibilitando a compilação de partes selecionadas do corpus (Machálek, 2020, p. 7004).

O KonText tem, ainda, a particularidade de em caso de crescimento da base de utilizadores, poder adicionar outro servidor para um melhor desempenho do sistema.

Não obstante, foram consultados outros projetos de interface, os quais não se encontram descritos neste capítulo, mas que merecem ser mencionados dada a sua pertinência e relevância no contexto desta dissertação. Alguns desses exemplos são o projeto OPUS[75] (Tiedemann & Nygaard, 2004), KorAP[76] (Diewald et al., 2019), TEITOK[77] (Janssen, 2016), Procura-PALavras[78] (Soares et al., 2018) e PEAPL2_PLE[79] (Martins, 2013).

Estas interfaces, à semelhança das acima apresentadas, reúnem uma série de características e funcionalidades, que se relacionam com este projeto e contribuíram para a reflexão e criação da prova-conceito que se descreve na secção 1.3 sobre o protótipo de interface de pesquisa.

6.3. ARQUITETURA DO PROTÓTIPO DE INTERFACE DE PESQUISA

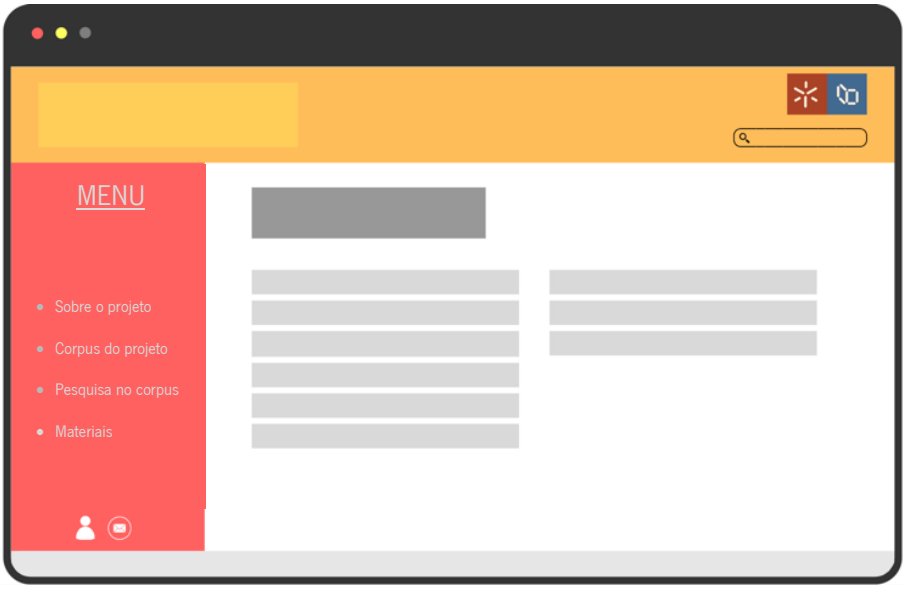
Tendo em consideração os projetos acima mencionados, onde foi possível identificar, de forma sucinta, algumas das suas características principais, apresenta-se nesta secção a prova-conceito da proposta de interface de pesquisa.

Pretende-se que a interface seja funcional e de uso intuitivo para o utilizador. Através da ilustração 16[80], observa-se uma representação gráfica do protótipo da página inicial da interface de pesquisa.

[75] https://opus.nlpl.eu/
[76] https://korap.ids-mannheim.de/
[77] http://www.teitok.org/index.php?action=home
[78] http://p-pal.di.uminho.pt/about/project
[79] https://teitok2.iltec.pt/peapl2-ple/index.php?action=home
[80] A ilustração que se apresenta serve apenas para acompanhar o texto e fornecer ao leitor uma componente gráfica meramente sugestiva daquilo que se pretende para a criação do protótipo de interface de pesquisa descrito nesta dissertação.

Ilustração 16

Apresentação da página inicial do protótipo de interface de pesquisa.



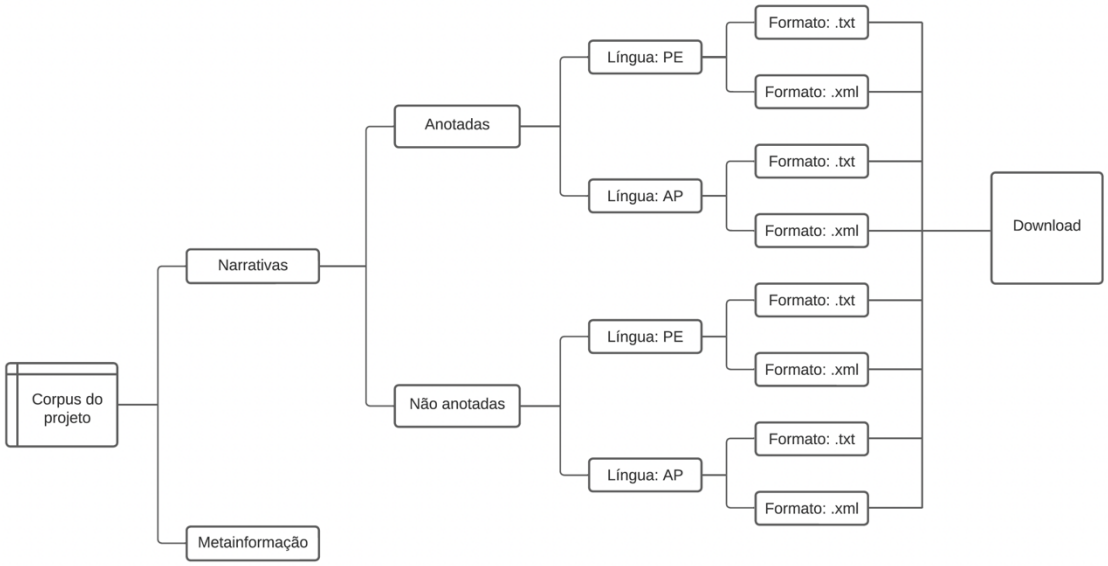
Tendo como elemento sugestivo a ilustração 16, observa-se que o menu principal se encontra numa barra lateral do lado esquerdo da interface (zona a vermelho), por se considerar, a nível visual, ser uma área mais intuitiva para o utilizador e que, assim, este tenha acesso imediato aos conteúdos que poderá aceder através desta interface de pesquisa. Os itens que constam no menu são "Sobre o projeto", "Corpus do projeto", "Pesquisa no corpus" e "Materiais".

No primeiro item designado "Sobre o projeto", pretende-se que o utilizador, ao aceder, possa ter uma apresentação sobre a interface de pesquisa, objetivos e os materiais disponíveis para, assim, dar início à utilização dos materiais e recursos na interface.

Por sua vez, no item "Corpus do projeto", o utilizador poderá ter acesso aos ficheiros corpus para *download* nos formatos .txt e .xml, em ambas versões (anotados e não anotados), assim como a metainformação recolhida através dos questionários sociolinguísticos. Atente-se na figura 16, onde é possível visualizar a arquitetura do item "Corpus do projeto".

Figura 16

Arquitetura do item "Corpus do projeto".

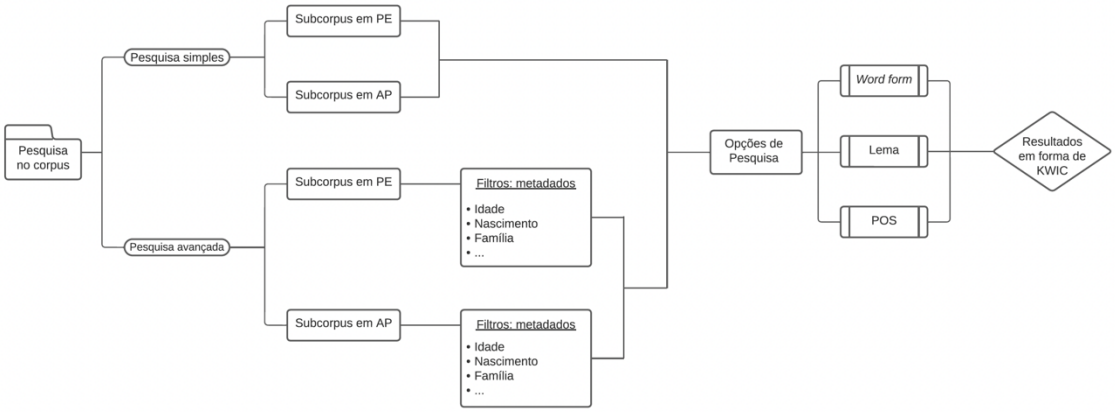


Ao aceder ao item "Corpus do projeto", o utilizador poderá escolher se quer ter acesso às narrativas ou à metainformação. Neste item, a metainformação, ao ser selecionada, aparecerá apenas para consulta. Caso o utilizador pretenda aceder aos ficheiros do corpus, deverá selecionar a opção "Narrativas" e depois escolher a versão que pretende obter: anotada ou não anotada. Dependendo da versão que o utilizador escolher, deverá, depois, selecionar a língua à qual pretende ter acesso – se às narrativas em PE ou em AP. Por último, deverá escolher o formato do ficheiro para *download*, estando as narrativas disponíveis em formato .txt e .xml. Depois de descarregados os ficheiros do corpus, o utilizador poderá visualizar os dados e proceder a uma análise ou utilizá-los para serem processados por uma ferramenta externa a outros níveis.

O terceiro item, "Pesquisa no corpus", permitirá que o utilizador efetue pesquisas morfossintáticas no corpus. Atente-se na figura 17 que mostra a arquitetura do item "Pesquisa no corpus".

Figura 17

Arquitetura do item "Pesquisa no corpus".



Ao aceder a este item, o utilizador poderá efetuar pesquisas no corpus e optar se pretende que a sua pesquisa seja 'simples' ou 'avançada'. Se optar por uma pesquisa simples, o utilizador poderá efetuar a sua pesquisa com base em todos os textos do subcorpus em PE ou em AP, dependendo da língua escolhida. Por conseguinte, poderá, depois, efetuar pesquisas aos níveis da *word form*, lemas ou POS. Os resultados aparecerão em forma de KWIC. Por outro lado, o utilizador poderá proceder a uma pesquisa mais personalizada, se optar pela pesquisa 'avançada'. Assim, depois de escolher o subcorpus onde pretende efetuar a sua pesquisa, poderá selecionar alguns filtros referentes à metainformação de cada narrativa para que a sua pesquisa se restrinja apenas aos textos com as características que o utilizador selecionar. Os filtros incidem sobre a informação da criança, podendo o utilizador selecionar, por exemplo, a idade, o lugar de nascimento (se o informante nasceu em Portugal ou na Suíça), há quanto tempo frequenta aulas de PLH, entre outros filtros que reflitam o entorno e comportamento da criança no que concerne à sua exposição ao PE e/ou AP.

Após a pesquisa, o utilizador poderá visualizar e ter acesso a todas as informações sobre a palavra pesquisada, nomeadamente os vários contextos de utilização, quantas vezes ocorre, a classe de palavra atribuída e o lema correspondente. Os resultados aparecerão em formato KWIC. Desta forma, será possível observar o comportamento da palavra pesquisada mediante os vários contextos em que se insere, bem como a classe ou classes gramaticais etiquetadas e respetivo lema.

Por último, o item "Materiais" pretende ser uma secção da interface para alojar dados linguísticos recolhidos, materiais de apoios como artigo científicos, manuais para a utilização de ferramentas de processamento de texto ou documentos para a recolha de dados linguísticos e sociolinguísticos. Desta forma, estes materiais estarão acessíveis a toda a comunidade e poderão servir de objeto de estudo para outros projetos de investigação, contribuindo, assim, para a sustentabilidade, preservação e reaproveitamento dos recursos linguísticos.

A criação deste protótipo de interface de pesquisa pretende democratizar a informação e tornar acessível os dados linguísticos compilados para este projeto de dissertação, bem como de outros projetos e, assim, contribuir para a sua reutilização e sustentabilidade. Por outro lado, representa uma forma de contribuir para a criação de materiais e recursos em PE e AP, tanto a nível de corpora como ao nível de ferramentas que disponibilizem o acesso a dados recolhidos neste âmbito de investigação. Em consequência, será, também, uma maneira de expandir e fomentar a investigação linguística com falantes de herança, dado que cada vez mais crianças experienciam a aquisição da linguagem em contexto migratório.

CONSIDERAÇÕES FINAIS

Baseado num corpus de aprendizagem, a presente dissertação centrou-se no processamento linguístico de narrativas, que com recurso a técnicas de PLN, possibilitaram a lematização e a etiquetação morfossintática dos textos, através da ferramenta *Sketch Engine*. Por se tratar de um corpus de aprendizagem, rico em dados linguísticos, foi possível identificar várias características linguísticas, que refletem desde a proficiência às particularidades da produção linguística dos informantes em ambas línguas. Para além disso, e como já fora mencionado, a criação de um corpus eletrónico permite não só a anotação para a extração de dados linguísticos, como a preservação das narrativas, oferecendo uma visão dos padrões linguísticos únicos deste grupo particular de falantes.

Não obstante, é importante mencionar alguns dos obstáculos encontrados durante o processamento do corpus. Os *corpora* de aprendizagem constituem um dos conjuntos de dados linguísticos que mais resistência oferecem durante o seu tratamento, isto porque, como o próprio nome indica, compilam dados de informantes que estão a adquirir ou adquiriram uma segunda língua na infância e uma língua de herança. Por conseguinte, embora sejam bastante ricos e diversificados, requerem um tratamento especial, dado que a probabilidade de que estes contenham erros ortográficos ou palavras e/ou expressões fruto do contacto linguístico (*code-switching*) é extremadamente elevada. Para além desse fator a ter em consideração, também as ferramentas disponíveis para o processamento dos *corpora* têm como referências *corpora* onde as línguas respeitam todas as formas estruturais e gramaticais exigidas. Assim, quando um tokenizador ou etiquetador se confronta com este tipo de dados linguísticos, ao não reconhecer os erros, poderá ter dificuldade em identificar ou etiquetar determinadas palavras do corpus em análise. Desta forma, os dados obtidos através de um corpus de aprendizagem, por se tratar de dados com características especiais, e as ferramentas disponíveis para o processamento de texto, que são treinadas com textos retirados de fontes onde a língua obedece às normas estipuladas, foram os dois principais obstáculos encontrados. Por esse motivo, procedeu-se à pesquisa e testagem de várias ferramentas utilizadas para o processamento de corpora, com o objetivo de encontrar a que melhor se adequava ao corpus em estudo nesta dissertação e aos objetivos da mesma.

Sem a necessidade de conhecimentos informáticos prévios, o *Sketch Engine* possibilitou o processamento das narrativas de forma intuitiva e bastante gráfica, tendo sido a ferramenta eleita dado à sua sensibilidade para questões no âmbito da linguística. Devido às suas funcionalidades, esta ferramenta possibilitou a exploração dos resultados através de linhas de concordância e listas de frequência, que possibilitaram, ainda, a criação de gráficos para uma melhor visualização e perceção dos resultados. Estas foram apresentadas ao longo da dissertação, tendo sempre em consideração os erros ortográficos presentes, que inevitavelmente, influenciaram alguns resultados. Não obstante, os resultados obtidos permitiram que se identificassem, por exemplo, os *tokens* e lemas mais e menos frequentes no corpus, assim como as classes de palavras, que serviram de base para a discussão.

Dado o contexto, importa reconhecer uma certa fragilidade que os resultados obtidos após o processamento do corpus possam ter, mas que, por outro lado, é importante sublinhar a riqueza que estes elementos conferem ao estudo. Assim, e apesar de na maioria dos casos se proceder a uma normalização ortográfica para que o corpus não crie atritos com a ferramenta de processamento, a retenção de ditos elementos torna-se benéfica, por exemplo, para a compreensão do desenvolvimento linguístico, do comportamento do aprendente e identificação de padrões e estratégias de aprendizagem, assim como na realização de estudos comparativos.

Também a recolha e inclusão dos dados extralinguísticos, como a metainformação recolhida através dos questionários sociolinguísticos, constitui, no estudo, um fator de extrema relevância, dado que possibilita o conhecimento das variáveis que influenciam o processo de aquisição de linguagem por parte dos informantes. Estas variáveis podem ir desde os dados biométricos do informante e familiares aos dados demográficos dos mesmos, que permitem uma melhor compreensão e interpretação dos dados recolhidos por parte dos investigadores. Assim, este conjunto de informação permite o cruzamento de variáveis, que possibilitam tecer conclusões de maneira mais assertiva.

Por seu turno, a apresentação da prova-conceito do protótipo de interface de pesquisa constitui a segunda parte deste projeto de dissertação, que por um lado pretende contribuir para melhorar a acessibilidade e utilização deste e outros recursos linguísticos, através do armazenamento de dados nas suas versões anotada e não anotada. Por outro lado, esta interface pretende fornecer um conjunto de funcionalidades que permitem o utilizador efetuar pesquisas para explorar os

dados armazenados. Tomando partido dos meios digitais disponíveis atualmente, esta plataforma contribuirá, ainda, com a preservação e a utilização sustentável deste tipo de recursos linguísticos.

Em suma, esta dissertação não só sublinha a interação linguística dinâmica no bilinguismo de herança, como também sublinha o potencial da análise de corpus para desvendar padrões únicos de utilização do PE e do AP num contexto migratório. Espera-se que este esforço interdisciplinar inspire outras investigações e avanços na compreensão do bilinguismo de herança no contexto mais alargado da exploração linguística, contribuindo para o estreitamento das relações luso-alemãs. As narrativas produzidas por estas crianças lusodescendentes são um testemunho da riqueza da diversidade cultural e linguística, que remontam para a notável diáspora portuguesa, particularmente presente em países como a França, Luxemburgo, Suíça e Alemanha, que mantém ligações com a sua herança portuguesa ao mesmo tempo que abraçam as culturas dos seus países de acolhimento.

REFERÊNCIAS BIBLIOGRÁFICAS

Abeillé, A. (2003). *Treebanks: Building and Using Parsed Corpora* (Vol. 20). Springer Netherlands. https://doi.org/10.1007/978-94-010-0201-1

Ädel, A. (2020). Corpus Compilation. In M. Paquot & S. T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 3–24). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_1

Adolphs, S. (2006). *Introducing Electronic Text Analysis*. Routledge. https://doi.org/10.4324/9780203087701

Afiah, N. (2020). Error Analysis of The Spoken English of The First Semester Students of Arabic Education of IAI DDI Polewali Mandar. *Loghat Arabi: Jurnal Bahasa Arab Dan Pendidikan Bahasa Arab*, *1*(1), 37–48. https://doi.org/10.36915/la.v1i1.4

Albi, A. B. (2019). How corpora can assist legal translation learners: The GENTT TransTools Corpora platform and Sketch Engine. *Quaderns de Filologia - Estudis Lingüístics XXIV*, *24*(24), 21–38. https://doi.org/10.7203/qf.24.16297

Almeida, L., & Flores, C. (2017). Bilinguismo. In M. J. Freitas & A. L. Santos (Eds.), *Aquisição de língua materna e não materna: Questões gerais e dados do português* (pp. 275–304). Language Science Press. https://doi.org/10.5281/zenodo.889439

Alyafeai, Z., Al-shaibani, M. S., Ghaleb, M., & Ahmad, I. (2023). Evaluating Various Tokenizers for Arabic Text Classification. *Neural Processing Letters*, *55*(3), 2911–2933. https://doi.org/10.1007/s11063-022-10990-8

Anthony, L. (2004). *AntConc*. Waseda University.

Anthony, L. (2009). Issues in the Design and Development of Software Tools for Corpus Studies: The Case for Collaboration. In P. Baker (Ed.), *Contemporary Corpus Linguistics* (pp. 87–104). Continuum Press.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, *30*(2), 141–161. https://doi.org/10.17250/khisli.30.2.201308.001

Anthony, L. (2015). *TagAnt (Version 1.2.0)*. Waseda University.

Antunes, S., Mendes, A., Gonçalves, A., Janssen, M., Alexandre, N., Avelar, A., Castelo, A., Duarte, I., Freitas, M. J., Pascoal, J., & Pinto, J. (2016). Apresentação do Corpus de Português Língua Estrangeira/Língua Segunda – COPLE2. *Revista Da Associação Portuguesa de Linguística*, *1*, 37–56. https://doi.org/10.21747/2183-9077/rapla3

Araújo, S., & Trabulo, P. (2014). Da linguística de corpus ao ensino/aprendizagem de línguas: da teoria à prática. *Revista de Letras, II*, *13*, 7–21.

Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., & Yarowsky, D. (Eds.). (1999). *Natural Language Processing Using Very Large Corpora* (Vol. 11). Springer Netherlands. https://doi.org/10.1007/978-94-017-2390-9

Baayen, R. H. (2001). Word Frequencies. In *Word Frequency Distributions. Text, Speech and Language Technology* (Vol. 18, pp. 1–38). Springer. https://doi.org/10.1007/978-94-010-0844-0_1

Baker, P. (2006). *Using Corpora in Discourse Analysis* (1st ed.). Bloomsbury Publishing Plc.

Barbaresi, A. (2018). A corpus of German political speeches from the 21st century. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 792–797. http://www.spiegel.de/wissenschaft/mensch/afd-auf-eure-werte-

Barlow, M. (2000). *MonoConc Pro 2.0*. Athelstan.

Batoréo, H. J. (2000). *Expressão do Espaço no Português Europeu. Contributo Psicolinguístico para o Estudo da Linguagem e Cognição* [Fundação Calouste Gulbenkian e Fundação para a Ciência e a Tecnologia, Ministério da Ciência e da Tecnologia]. https://doi.org/10.21415/T5701M

Benmamoun, E., Montrul, S., & Polinsky, M. (2013). Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics*, *39*(3–4), 129–181. https://doi.org/10.1515/tl-2013-0009

Bennett-Kastor, T. (2002). The "frog story" narratives of Irish–English bilinguals. *Bilingualism: Language and Cognition*, *5*(2), 131–146. https://doi.org/10.1017/s1366728902000238

Bhatia, T. K. (2006). Bilingualism and Second Language Learning. In *Encyclopedia of Language & Linguistics* (pp. 16–22). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00620-9

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Bloomfield, L. (1933). *Language*. Holt, Rinehart & Winston.

Blumer, H. (1930). Review of J. Piaget's The Child's Conception of the World. *American Journal of Sociology*, *36*(1), 150–151.

Bonial, C., Hargraves, O., & Palmer, M. (2013). *Expanding VerbNet with Sketch Engine*.

Boulton, A. (2017). Data-Driven Learning and Language Pedagogy. In S. L. Thorne & S. May (Eds.), *Language, Education and Technology* (pp. 181–192). Springer International Publishing. https://doi.org/10.1007/978-3-319-02237-6_15

Brants, T. (2006). Part-of-Speech Tagging. *Part-of-Speech Tagging*, 221–230.

Bruner, J. (1981). The Social Context of Language Acquisition. *Language & Communication*, *1*(2/3), 155–178.

Buchweitz, A., & Prat, C. (2013). The bilingual brain: Flexibility and control in the human cortex. *Physics of Life Reviews*, *10*(4), 428–443. https://doi.org/10.1016/j.plrev.2013.07.020

Burnard, L. (2009). *What is the BNC?* Http://Www.Natcorp.Ox.Ac.Uk/Corpus/Index.Xml?ID=intro.

Cardoso, A., Magro, C., Braz, J., & Nunes, T. (2014). *CUTe: Corpus of Portuguese Undergraduates' Texts Um recurso para a investigação em escrita académica em português*.

Chahine, I. K., & Uetova, E. (2023). Spelling issues: what learner corpora can reveal about L2 orthography. *Corpus [Online]*, *24*, 1–19. https://doi.org/10.4000/corpus.8226

CHILDES. (2000). *CHILDES Portuguese Batoréo Corpus*.

Chomsky, N. (1959). Review of B. F. Skinner's Verbal Behavior. *Language*, *35*(1), 26–58. https://doi.org/10.4159/harvard.9780674594623.c6

CLUL. (2019a). *Acquisition of European Portuguese Databank*. Centro de Linguística Da Universidade de Lisboa.

CLUL. (2019b). *Recolha de dados de PLE*. Https://Www.Clul.Ulisboa.Pt/Recurso/Recolha-de-Dados-de-Ple.

Côrrea, L. M. S. (2018). Conciliando processamento linguístico e teoria de língua no estudo da aquisição da linguagem. In L. M. S. Corrêa (Ed.), *Aquisição da Linguagem e Problemas do Desenvolvimento Linguístico* (2nd ed.). Editora PUC-Rio.

Correia, L., & Flores, C. (2021). Questionário sociolinguístico parental para famílias emigrantes bilingues (QuesFEB) uma ferramenta de recolha de dados sociolinguísticos de crianças falantes de herança. *Linguística: Revista de Estudos Linguísticos Da Universidade Do Porto*, *16*, 75–102. https://doi.org/10.21747/16466195/ling16a3

Coughlin, J. (1990). Perspectives on Natural Language Processing. In *Source: The French Review* (Vol. 64, Issue 1). http://www.jstor.orgURL:http://www.jstor.org/stable/395699http://www.jstor.org/stable/395699?seq=1&cid=pdf-reference#references_tab_contents

Crystal, D. (2003). *English as a Global Language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486999

Cunha, C., & Sintra, L. (2017). *Nova gramática do português contemporâneo [recurso eletrónico]* (7th ed.). Lexicon.

Dash, N. S. (2021). Language Corpora Annotation and Processing. In *Language Corpora Annotation and Processing*. Springer Singapore. https://doi.org/10.1007/978-981-16-2960-0

Dash, N. S., & Arulmozi, S. (2018). History, features, and typology of language corpora. In *History, Features, and Typology of Language Corpora*. Springer Singapore. https://doi.org/10.1007/978-981-10-7458-5

Davies, M. (2013). *corpus.byu.edu* .

Dehé, N. (2015). Particle verbs in Germanic. In P. O. Müller, I. Ohnheiser, S. Olsen, & F. Rainer (Eds.), *Word Formation. An International Handbook of the Languages of Europe* (pp. 611–626). De Gruyter. https://doi.org/10.1515/9783110246254-037

Del Río, I., Antunes, S., Mendes, A., & Janssen, M. (2016). Towards error annotation in a learner corpus of Portuguese. *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 8–17. http://www.clul.ul.pt/en/research-teams/547

Diewald, N., Mititelu, V. B., & Kupietz, M. (2019). THE KORAP USER INTERFACE. ACCESSING COROLA VIA KORAP. *Revue Roumaine de Linguistique*, *64*(3), 265–277.

Durrant, P., & Siyanova-Chanturia, A. (2015). Learner corpora and psycholinguistics. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 1–748). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Duruttya, M. (2022). Corpus-based Linguistic Analysis of Business English Report Writing Papers by L2 English Language Speakers. *MUNI Journals*, *12*(2), 128–144.

Elsen, H. (2000). The acquisition of verb morphology. In M. Beers, B. Bogaerde, G. Bol, J. Jong, & C. Rooijmans (Eds.), *From Sound to Sentence: Studies on First Language Acquisition* (pp. 31–41). Centre for Language and Cognition Groningen.

Evison, J. (2010). What are the basics of analysing a corpus? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 122–135). Routledge.

Falchi, S. (2021). Islands Isles Isole A Sketch Engine Based Corpus Analysis. *Rhesis. International Journal of Linguistics, Philology and Literature*, *12*(1), 17–34.

Fenlon, J., & Hochgesang, J. A. (2022a). Introduction to Signed Language Corpora. In J. Fenlon & J. A. Hochgesang (Eds.), *Signed Language Corpora* (Vol. 25). Gallaudet University Press.

Fenlon, J., & Hochgesang, J. A. (2022b). *Signed Language Corpora*. https://doi.org/https://doi.org/10.2307/j.ctv2rcnfhc

Fernández-Dobao, A., & Herschensohn, J. (2021). Acquisition of Spanish verbal morphology by child bilinguals: Overregularization by heritage speakers and second language learners. *Bilingualism: Language and Cognition*, *24*(1), 56–68. https://doi.org/10.1017/S1366728920000310

Ferreira, T., Santos, I., Carapinha, C., Martins, C., Pereira, I., Rio-Torto, G., Inverno, L., Pereira, R., Ferreira, C., Sousa, S., & Chapouto, S. (2023). Construção do corpus "Produção Oral em Provas de Português L2" (POPL2). *Études Romanes de Brno*, *1*, 245–261. https://doi.org/10.5817/ERB2023-1-14

Fiestas, C. E., & Peña, E. D. (2004). Narrative Discourse in Bilingual Children. *Language, Speech, and Hearing Services in Schools*, *35*(2), 155–168. https://doi.org/10.1044/0161-1461(2004/016)

Flores, C. (2019a). BILINGUISMO INFANTIL. UM LEGADO VALIOSO DO FENÓMENO MIGRATÓRIO. *Diacrítica*, *31*(3), 237–250. https://doi.org/10.21814/diacritica.395

Flores, C. (2019b). BILINGUISMO INFANTIL. UM LEGADO VALIOSO DO FENÓMENO MIGRATÓRIO. *Diacrítica*, *31*(3), 237–250. https://doi.org/10.21814/diacritica.395

Flores, C., Gonçalves, M. L., Rinke, E., & Torregrossa, J. (2022). Perspetivas múltiplas sobre a competência bilingue de crianças lusodescendentes residentes na Suíça. *Revista Portuguesa de Educação*, *35*(1), 102–131. https://doi.org/10.21814/rpe.24205

Flores, C., & Melo-Pfeifer, S. (2014). O conceito "Língua de Herança" na perspetiva da Linguística e da Didática de Línguas: considerações pluridisciplinares em torno do perfil linguístico das crianças lusodescendentes na Alemanha. *DOMÍNIOS DE LINGU@GEM*, *8*(3), 16–45. http://www.seer.ufu.br/index.php/dominiosdelinguagem

Flores, C., Rinke, E., Torregrossa, J., & Weingärtner, D. (2022). Language separation and stable syntactic knowledge: verbs and verb phrases in bilingual children's narratives. *Journal of Portuguese Linguistics*, *5*(1), 1–30. https://doi.org/10.16995/jpl.8043

Francis, W. N. (1992). [Review of the book The London-Lund Corpus of Spoken English: Description and research Edited by Jan Svartvik]. *Language*, *68*(1), 196–199. https://doi.org/10.1353/lan.1992.0072

Freitas, C. (2015). Corpus, Linguística Computacional e as Humanidades Digitais. In M. S. Leite & C. Gabriel (Eds.), *Linguagem, Discurso, Pesquisa e Educação* (pp. 18–46). DP et alii.

Freitas, M. J. (1997). *Aquisição da Estrutura Silábica do Português Europeu*. University of Lisbon.

Fromm, G. (2003). O USO DE CORPORA NA ANÁLISE LINGÜÍSTICA 1. *Revista Factus*, *1*(1), 69–76.

Gagarina, N., Klop, D., Tsimpli, I. M., & Walters, J. (2016). Narrative abilities in bilingual children. In *Applied Psycholinguistics* (Vol. 37, Issue 1, pp. 11–17). Cambridge University Press. https://doi.org/10.1017/S0142716415000399

Gamallo, P., & Garcia, M. (2017). LinguaKit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática*, *9*(1), 19–28. https://doi.org/10.21814/lm.9.1.243

Généreux, M., Hendrickx, I., & Mendes, A. (2012). A large Portuguese corpus on-line: cleaning and preprocessing. In H. et al Caseli (Ed.), *Proceedings of the 10th International Conference PROPOR1012* (pp. 113–120). Springer-Verlag. http://hdl.handle.net/10451/37430

Genesee, F. (2001). Bilingual first language acquisition: exploring the limits of the language faculty. *Annual Review of Applied Linguistics*, *21*, 153–168. https://doi.org/10.1017/S0267190501000095

Geyken, A., & Kupietz, M. (2016). Editorial. *Journal for Language Technology and Computational Linguistics (JLTCL)*, *31*(1).

Gibson, M., & Ruotolo, C. (2003). Beyond the Web: TEI, the Digital Library, and the Ebook Revolution. *Computers and the Humanities*, *37*(1), 57–63. https://doi.org/10.1023/A:1021895322291

Gilquin, G. (2020). Learner Corpora. In M. Paquot & S. T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 283–303). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_13

Glória, Y. A. L., Hanauer, L. P., Wiethan, F. M., Nóro, L. A., & Mota, H. B. (2016). O uso das conjunções por crianças com desenvolvimento típico de linguagem. *CoDAS*, *28*(3), 221–225. https://doi.org/10.1590/2317-1782/20162015107

Goldfarb, C. F. (1990). *The SGML Handbook* (Y. Rubinsky, Ed.). Oxford University Press.

Granger, S. (2003). The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, *37*(3), 538. https://doi.org/10.2307/3588404

Granger, S. (2008). Learner Corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (Vol. 1, pp. 259–275). Walter de 5 Gruyter. https://doi.org/10.1002/9781405198431.wbeal0669

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: learner corpus research – past, present and future. In *The Cambridge Handbook of Learner Corpus Research* (pp. 1–6). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.001

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (Vol. 6). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.6

Granja, M. Á. de la, & Romero, M. N. (2015). O processo de lematização no Tesouro do léxico patrimonial galego e português. In F. C. Altino, G. A. L. Milani, & R. E. S. B. Rodrigues (Eds.), *Anais do III CIDS: Congresso Internacional de Dialetologia e Sociolinguística* (pp. 848–862). Universidade Estadual de Londrina.

Greenberg, J. (1998). The Applicability of Natural Language Processing (NLP) to Archival Properties and. In *Source: The American Archivist* (Vol. 61, Issue 2).

Gries, S. Th. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, *3*(5), 1225–1241. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Gries, S. Th., & Berez, A. L. (2017). Linguistic Annotation in/for Corpus Linguistics. In *Handbook of Linguistic Annotation* (pp. 379–409). Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2_15

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3–15. https://doi.org/10.1016/0093-934X(89)90048-5

Gut, U. (2020). Spoken Corpora. In M. Paquot & S. Th. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 235–256). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_11

Hansen, B. (2018). *Corpus Linguistics and Sociolinguistics*. BRILL. https://doi.org/10.1163/9789004381520

Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Herdeiro, A. J., & Barbosa, P. (2015). O FENÓMENO DO QUEÍSMO NO FALAR BRACARENSE: UM ESTUDO SOCIOLINGUÍSTICO "QUEÍSMO" IN THE SPEECH OF BRAGA: A SOCIOLINGUISTICS STUDY. *Diacrítica - Ciências Da Linguagem*, *29*(1), 327–351.

Higginson, R. (1990). An Update on The Childes/Bib (formerly Isu/Childes) Database. *Journal of Child Language*, *17*(2), 473–479. https://doi.org/10.1017/S0305000900013878

Hoff-Ginsberg, E., & Shatz, M. (1982). Linguistic input and the child's acquisition of language. *Psychological Bulletin*, *92*(1), 3–26. https://doi.org/10.1037/0033-2909.92.1.3

Hovy, E., & Lavid, J. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, *20*(1), 13–36.

Hu, C., & Yang, B. (2015). Using Sketch Engine to Investigate Synonymous Verbs. *International Journal of English Linguistics*, *5*(4), 29–41. https://doi.org/10.5539/ijel.v5n4p29

Huang, X., & Li, D. (2010). An Overview of Modern Speech Recognition. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed., pp. 339–366). Chapman & Hall/CRC.

Hunston, S. (2002a). *Corpora in Applied Linguistics* (S. Hunston, Ed.; 1st ed.). Cambridge University Press.

Hunston, S. (2002b). Methods in corpus linguistics: Interpreting concordance lines. In *Corpora in Applied Linguistics* (pp. 38–66). Cambridge University Press. https://doi.org/10.1017/CBO9781139524773.004

Hunston, S. (2022). *Corpora in Applied Linguistics* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108616218

Ingram, D. (1989). *First Language Acquisition: Method, Description and Explanation*. Cambridge University Press.

ISD. (2023). *Development and Maintenance of Contemporary Written Corpora*. Mannheim: Institut Für Deutsche Sprache.

Järvinen, T. (2003). Bank of English and Beyond. In A. Abeillé (Ed.), *Treebanks. Text, Speech and Language Technology* (Vol. 20, pp. 43–59). Springer. https://doi.org/10.1007/978-94-010-0201-1_3

Jentsch, P., & Porada, S. (2020). From Text to Data. In S. Schwandt (Ed.), *Digital Methods in the Humanities: Challenges, Ideas, Perspectives* (pp. 89–128). Bielefeld University Press. https://doi.org/10.2307/j.ctv2f9xskk.6

Joshi, A. K. (1991). *Natural Language Processing*. http://about.jstor.org/terms

Kapalková, S., Polišenská, K., Marková, L., & Fenton, J. (2016). Narrative abilities in early successive bilingual Slovak–English children: A cross-language comparison. *Applied Psycholinguistics*, *37*(1), 145–164. https://doi.org/10.1017/S0142716415000454

Kettunen, K. (2014). Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? *Journal of Quantitative Linguistics*, *21*(3), 223–245. https://doi.org/10.1080/09296174.2014.911506

Kharis, M., Kisyani, Suhartono, Pairin, U., & Darni. (2021). How to Lemmatize German Words with NLP-Spacy Lemmatizer? In M. Hidayati, Y. Basthomi, F. M. Ivone, N. Ariani, & A. Tohe (Eds.), *Proceedings of the International Seminar on Language, Education, and Culture (ISoLEC 2021)* (pp. 189–193). Atlantis Press. https://doi.org/10.2991/assehr.k.211212.036

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*, 7–36.

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of finnish text documents. *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, 625–633. https://doi.org/10.1145/1031171.1031285

Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Brown University Press.

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DEREKO: A Primordial Sample for Linguistic Research. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 1848–1854.

Kupietz, M., Lüngen, H., Kamocki, P., & Witt, A. (2018). The German Reference Corpus DeReKo: New Developments – New Opportunities. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 4353–4360.

Lan, G., Lucas, K., & Sun, Y. (2019). Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. *System*, *85*, 102–116. https://doi.org/10.1016/j.system.2019.102116

Lange, H. (2022). Metadata Formats for Learner Corpora: Case Study and Discussion. In D. Alfter, E. Volodina, T. François, P. Desmet, F. Cornillie, A. Jönsson, & E. Rennes (Eds.), *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning* (pp. 108–113). LiU Electronic Press. https://doi.org/10.3384/ecp190011

Leacock, C., Chodorow, M., & Tetreault, J. (2015). Automatic grammar - and spell-checking for language learners. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 1–748). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Leech, G. (1993). Corpus Annotation Schemes. *Literary and Linguistic Computing*, *8*(4), 275–281. http://llc.oxfordjournals.org/

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 1–18). Addison Wesley Longman.

Lenneberg, E. H. (1967). The Biological Foundations of Language. *Hospital Practice*, *2*(12), 59–67. https://doi.org/10.1080/21548331.1967.11707799

Liddy, E. D. (2001). *Natural Language Processing Natural Language Processing Natural Language Processing 1*. https://surface.syr.edu/istpub

Lorandi, A., Cruz, C. R., & Scherer, A. P. R. (2011). Aquisição da linguagem. *Verba Volant*, *2*(1), 144–166.

Lozano, C., Teixeira, J., & Madeira, A. (2021). Corpora and L2 acquisition: the L1 Portuguese – L2 Spanish subcorpus of CEDEL2. *Revista Da Associação Portuguesa de Linguística*, *8*, 121–136. https://doi.org/10.26334/2183-9077/rapln8ano2021a10

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 1–748). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Lynch, A. (2017). Bilingualism and Second Language Acquisition. In N. Van Deusen-Scholl & S. May (Eds.), *Second and Foreign Language Education. Encyclopedia of Language and Education* (3rd ed., Vol. 10, pp. 43–55). Springer International Publishing. https://doi.org/10.1007/978-3-319-02246-8_5

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 7003–7008.

MacWhinney, B. (1992). The CHILDES project: tools for analyzing talk. *Child Language Teaching and Therapy*, *8*(2), 217–218. https://doi.org/10.1177/026565909200800211

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods*, *51*(4), 1919–1927. https://doi.org/10.3758/s13428-018-1174-9

Maden-Weinberger, U. (2015). "Hätte, wäre, wenn..." A pseudo-longitudinal study of subjunctives in the Corpus of Learner German (CLEG). *International Journal of Learner Corpus Research*, *1*(1), 25–57. https://doi.org/10.1075/ijlcr.1.1.02mad

Marçalo, M. J. (2009). *Fundamentos para uma Gramática de Funções Aplicada ao Português* (Vol. 4). Centro de Estudos em Letras Universidade de Évora.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, *35*(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004

Martins, C. (2013). O Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2/CELGA). Caracterização e desenvolvimento de uma infra-estrutura de investigação. In R. Bizarro, M. A. Moreira, & C. Flores (Eds.), *Português Língua Não Materna: Investigação e Ensino* (pp. 69–80). Lidel.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics*, *39*, 74–92. https://doi.org/10.1017/S0267190519000096

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction* (2nd ed.). Edinburgh University Press.

McManus, K. (2021). Introducing Crosslinguistic Influence. In *Crosslinguistic Influence and Second Language Learning* (pp. 1–19). Routledge. https://doi.org/10.4324/9780429341663-1

Meir, N., & Janssen, B. (2021). Child Heritage Language Development: An Interplay Between Cross-Linguistic Influence and Language-External Factors. *Frontiers in Psychology*, *12*, 1–17. https://doi.org/10.3389/fpsyg.2021.651730

Meisel, J. M. (2011). *First and Second Language Acquisition: Parallels and Differences*. Cambridge University Press.

Mendes, A. (2016). Linguística de Corpus e outros usos dos corpora em linguística. In A. M. Martins & E. Carrilho (Eds.), *Manual de Linguística Portuguesa* (Vol. 16, pp. 224–251). Walter de Gruyter. http://hdl.handle.net/10451/30696

Mendes, A., Antunes, S., Janssen, M., & Gonçalves, A. (2016). The COPLE2 Corpus: a Learner Corpus for Portuguese. *Proceedings of the Tenth Language Resources and Evaluation Conference – LREC'16*, 3207–3214. http://www.clul.ul.pt/en/research-teams/547

Michelfeit, J., Pomikálek, J., & Suchomel, V. (2014). Text Tokenisation Using unitok. In A. Horák & P. Rychlý (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014* (pp. 71–75). NLP Consulting.

Miller, D. (2020). Analysing Frequency Lists. In M. Paquot & S. Th. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 77–97). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_4

Montrul, S. (2015). *The Acquisition of Heritage Languages*. Cambridge University Press. https://doi.org/10.1017/CBO9781139030502

Mukherjee, J., & Götz, S. (2015). Learner corpora and learning context. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 1–748). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Nascimento, B., Fernanda, M., Mendes, A., Antunes, S., & Pereira, L. (2014). The Reference Corpus of Contemporary Portuguese and related resources. In T. B. Sardinha & T. Ferreira (Eds.), *Working with Portuguese Corpora* (pp. 237–256). Bloomsbury Publishing.

Nasseri, M., & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, *47*(100511), 1–11. https://doi.org/10.1016/j.asw.2020.100511

Neunerdt, M., Trevisan, B., Reyer, M., & Mathar, R. (2013). Part-Of-Speech Tagging for Social Media Texts. In I. Gurevych, C. Biemann, & T. Zesch (Eds.), *Language Processing and Knowledge in the Web. Lecture Notes in Computer Science* (Vol. 8105, pp. 139–150). Springer. https://doi.org/10.1007/978-3-642-40722-2_15

Newman, J., & Cox, C. (2020). Corpus Annotation. In M. Paquot & S. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 25–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_2

Nygaard, L., Priestley, J., Nøklestad, A., & Johannessen, J. B. (2008). Glossa: A multilingual, multimodal, configurable user interface. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 617–622.

Odlin, T. (2005). CROSSLINGUISTIC INFLUENCE AND CONCEPTUAL TRANSFER: WHAT ARE THE CONCEPTS? *Annual Review of Applied Linguistics*, *25*, 3–25. https://doi.org/10.1017/S0267190505000012

Operstein, N. (2021). The Lexicon. In N. Operstein (Ed.), *The Lingua Franca* (pp. 134–172). Cambridge University Press. https://doi.org/10.1017/9781009000161.006

Othero, G. de Á., & Ayres, M. R. (2014). Anotação morfológica automática de corpus de língua falada: desafios ao Aelius. *Texto Livre: Linguagem e Tecnologia*, *7*(2), 44–60.

Palmer, D. (2010). Text Preprocessing. In N. Indurkhya & F. J. Damerau (Eds.), *The Handbook of Natural Language Processing* (2nd ed., pp. 9–30). CRC Press.

Panunzi, A., Picchi, E., & Moneglia, M. (2004). Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus: C-Oral-Rom Italian. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (pp. 563–566). European Language Resources Association (ELRA).

Paquot, M. (2023). *Learner corpora around the world*. Https://Uclouvain.Be/En/Research-Institutes/Ilc/Cecl/Learner-Corpora-around-the-World.Html.

Piotrowski, M. (2012). Natural Language Processing for Historical Texts. *Synthesis Lectures on Human Language Technologies*, *5*(2), 1–157. https://doi.org/10.2200/S00436ED1V01Y201207HLT017

Põldvere, N., Johansson, V., & Paradis, C. (2021). On The London–Lund Corpus 2: design, challenges and innovations. *English Language and Linguistics*, *25*(3), 459–483. https://doi.org/10.1017/S1360674321000186

Proisl, T., Dykes, N., Heinrich, P., Kabashi, B., Blombach, A., & Evert, S. (2020). EmpiriST Corpus 2.0: Adding Manual Normalization, Lemmatization and Semantic Tagging to a German Web and CMC Corpus. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6142–6148. https://github.com/fau-klue/

Puerta-Díaz, M., De Mira, B. S., Martínez-Ávila, D., Ovalle-Perandones, M. A., & Grácio, M. C. C. (2021). Natural language processing in information metric studies: An analysis of the articles indexed by the web of science (2000-2019). *Encontros Bibli*, *26*. https://doi.org/10.5007/1518-2924.2021.e76886

Rajaa, G. N. (2016). Crosslinguistic Influence during Second Language Acquisition: The Reasons behind. *Journal of Translation and Languages*, *15*(1), 201–214.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549. https://doi.org/10.1075/ijcl.13.4.06ray

Read, J. (2000). Comprehensive measures of vocabulary. In J. Read (Ed.), *Assessing Vocabulary* (pp. 188–221). Cambridge University Press. https://doi.org/10.1017/CBO9780511732942.008

Reppen, R., & Ide, N. (2004). The American National Corpus. *Journal of English Linguistics*, *32*(2), 105–113. https://doi.org/10.1177/0075424204264856

Roch, M., Florit, E., & Levorato, C. (2016). Narrative competence of Italian–English bilingual children between 5 and 7 years. *Applied Psycholinguistics*, *37*(1), 49–67. https://doi.org/10.1017/S0142716415000417

Rodrigues, M., & Teixeira, A. (2015). Data Gathering, Preparation and Enrichment. In *Advanced Applications of Natural Language Processing for Performing Information Extraction* (pp. 13–26). Springer, Cham. https://doi.org/10.1007/978-3-319-15563-0_2

Rodríguez, P. (2022). Review of "TEITOK, a visual solution for XML/TEI encoding: editing, annotating and hosting linguistic corpora." *RIDE*, *15*, 1–21.

Rogošić, G. D. (2019). Using Learner Corpus Evidence in Error Analysis. *10th International Language Conference on »The Importance of Learning Professional Foreign Languages for Communication between Cultures«*, 81–92. https://doi.org/10.18690/978-961-286-252-7.7

Rosen, A. (2014). *CzeSL-SGT-a corpus of non-native speakers' Czech with automatic annotation*. http://utkl.ff.cuni.cz/learncorp/andhttp://akces.ff.cuni.cz/.Thesitesincludebibliographylists.Formorehttp://utkl.ff.cuni.cz/~rosen/public/SeznamAutoChybR0R1_en.html.1

Rühlemann, C. (2014). Data, methods, and tools. In *Narrative in English Conversation: A Corpus Analysis of Storytelling* (pp. 40–75). Cambridge University Press. https://doi.org/10.1017/CBO9781139026987.004

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science* (Vol. 2276, pp. 1–15). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45715-1_1

Sampson, G., & McCarthy, D. (2004). *Corpus Linguistics* (G. Sampson & D. McCarthy, Eds.; 1st ed.). Continuum.

Santos, A. L. (2006). *Minimal Answers. Ellipsis, Syntax and Discourse in the Acquisition of European Portuguese*. Universidade de Lisboa.

Santos, A. L., Généreux, M., Cardoso, A., Agostinho, C., & Abalada, S. (2014). A corpus of European Portuguese child and child-directed speech. In *Proceedings of the 9th Conference on Language Resources and Evaluation – LREC 2014* (pp. 1488–1491). European Language Resources Association (ELRA). http://childes.psy.cmu.edu/.

Sardinha, T. B. (2004). *Lingüística de Corpus* (Vol. 1). Manole. https://doi.org/10.1590/S0102-44502004000200014

Sarmento, S. (2010). LINGUÍSTICA DE CORPUS: HISTÓRICO, METODOLOGIA, CAMPOS DE APLICAÇÃO. In *Revista Trama* (Vol. 6).

Savoy, J., & Gaussier, E. (2010). Information Retrieval. In N. Indurkhya & F. J. Damerau (Eds.), *The Handbook of Natural Language Processing* (pp. 455–484). CRC Press.

Schallert, O. (2020). A Note on Misplaced or Wrongly Attached "zu" in German. *Journal of Germanic Linguistics*, *32*(1), 43–82. https://doi.org/10.1017/S1470542719000138

Schembri, A., & Cormier, K. (2022). Signed Language Corpora: Future Directions. In E. Shaw & J. A. Hochgesang (Eds.), *Signed Language Corpora* (1st ed., Vol. 25, pp. 196–220). Gallaudet University Press.

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, 1–9.

Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an Application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, & D. Yarowsky (Eds.), *Natural Language Processing Using Very Large Corpora* (Vol. 11, pp. 13–25). Springer Netherlands. https://doi.org/10.1007/978-94-017-2390-9

Schmidt, T. (2016). Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. *International Journal of Corpus Linguistics*, *21*(3), 396–418. https://doi.org/10.1075/ijcl.21.3.05sch

Schreibman, S. (2002). Computer-mediated Texts and Textuality: Theory and Practice. *Computers and the Humanities*, *36*(3), 283–293. https://doi.org/10.1023/A:1016178200469

Scott, M. (1996). *Wordsmith Tools* (1). Oxford University Press.

Silva, A. S. da, & Batoréo, H. (2010). Gramática cognitiva: estruturação conceptual, arquitectura e aplicações. In A. M. Brito (Ed.), *Gramática: História, Teorias, Aplicações* (pp. 229–251). Faculdade de Letras da Universidade do Porto.

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

Sinclair, J. (2004). *Trust the Text: Language, Corpus, and Discourse*. Routledge.

Skinner, B. F. (1957). *Verbal Behavior*. Appleton-Century-Crofts.

Soares, A. P., Iriarte, Á., de Almeida, J. J., Simões, A., Costa, A., Machado, J., França, P., Comesaña, M., Rauber, A., Rato, A., & Perea, M. (2018). Procura-PALavras (P-PAL): A Web-based interface for a new European Portuguese lexical database. *Behavior Research Methods*, *50*(4), 1461–1481. https://doi.org/10.3758/s13428-018-1058-z

Soehn, J.-P., Zinsmeister, H., & Rehm, G. (2008). Requirements of a user-friendly, general-purpose corpus query interface. In L. Burnard, K. Choukri, G. Rehm, T. Schmidt, & A. Witt (Eds.), *Proceedings of the LREC 2008 Workshop "Sustainability of Language Resources and Tools for Natural Language Processing"* (pp. 27–32).

Sokolov, J. L., & Snow, C. E. (1997). *Linguistic Society of America Handbook of Research in Language Development Using CHILDES* (Vol. 73, Issue 3).

Sperberg-McQueen, C. M. (1994). The Text Encoding Initiative. In A. Zampolli, N. Calzolari, & M. Palmer (Eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker* (pp. 409–427). Springer Netherlands. https://doi.org/10.1007/978-0-585-35958-8_21

Stocker, P. (2012). *A Student Grammar of German*. Cambridge University Press. https://doi.org/10.1017/CBO9781139028783

Stoll, S., & Schikowski, R. (2020). Child-Language Corpora. In M. Paquot & S. T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 305–327). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_14

Svartvik, J. (1990). *The London–Lund corpus of spoken English: Description and research* (Vol. 82). Lund University Press.

Taylor, L., & Barker, F. (2008). Using Corpora for Language Assessment. In N. H. Hornberger (Ed.), *Encyclopedia of Language and Education* (pp. 2377–2390). Springer US. https://doi.org/10.1007/978-0-387-30424-3_179

Tello, J. C. (2021). Grammatical, Lexical, Semantic, and Textual Annotation. In *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning* (pp. 179–190). Bielefeld University Press. https://doi.org/10.2307/j.ctv2f9xsw7.11

Torrijos, M. del M. R. (2009). EFFECTS OF CROSS-LINGUISTIC INFLUENCES ON SECOND LANGUAGE ACQUISITION: A CORPUS-BASED STUDY OF SEMANTIC TRANSFER IN WRITTEN PRODUCTION. *Revista de Lingüística y Lenguas Aplicadas*, *4*(1), 147–159. https://doi.org/10.4995/rlyla.2009.741

Tsimpli, I. M., Peristeri, E., & Andreou, M. (2016). Narrative production in monolingual and bilingual children with specific language impairment. *Applied Psycholinguistics*, *37*(1), 195–216. https://doi.org/10.1017/S0142716415000478

van Rooy, B. (2015). Annotating learner corpora. In *The Cambridge Handbook of Learner Corpus Research* (pp. 79–106). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.005

Vanhoutte, E. (2004). An Introduction to the TEI and the TEI Consortium. *Literary and Linguistic Computing*, *19*(1), 9–16. https://doi.org/10.1093/llc/19.1.9

Walker, D. E. (1994). The Ecology of Language. In A. Zampolli, N. Calzolari, & M. Palmer (Eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker* (pp. 359–375). Springer Netherlands. https://doi.org/10.1007/978-0-585-35958-8_19

Wallis, S. (2021). What might a corpus of parsed spoken data tell us about language? In *Statistics in Corpus Linguistics Research: A New Approach* (Vol. 1, pp. 1–14).

Weissenborn, J. (1989). A DATA BASE FOR THE STUDY OF FIRST LANGUAGE ACQUISITION. CHILDES: Child Language Data Exchange System. In *Language Sciences* (Vol. 11, Issue 2, pp. 259–265).

Westpfahl, S. (2014). STTS 2.0? Improving the tagset for the part-of-speech-tagging of German spoken data. *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, 1–10.

Wiese, H. (2011). Ein neuer urbaner Dialekt im multiethnischen Raum : Kiezdeutsch. In M. Messling, D. Läpple, & J. Trabant (Eds.), *Stadt und Urbanität: transdisziplinäre Perspektiven* (pp. 146–161). Kadmos.

Wiese, H., Freywald, U., Schalowski, S., Mayr, K., Daten, S., & Aus Urbanen Wohngebieten, J. (2012). Das KiezDeutsch-Korpus. *Deutsche Sprache*, *2*, 97–123. https://doi.org/https://doi.org/10.37307/j.1868-775X.2012.02.02

Williamson, G. (2009). *Type-Token Ratio of Written Language*. www.sltinfo.com

Wolters, A. P., & Kim, Y. G. (2023). Crosslinguistic influence on spelling in written compositions: Evidence from English-Spanish dual language learners in primary grades. *Reading and Writing*, 1–20. https://doi.org/10.1007/s11145-023-10416-4

Wulff, S., & Baker, P. (2020). Analyzing Concordances. In M. Paquot & S. Th. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 161–179). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_8

Wynne, M. (2004). *Developing Linguistic Corpora: a Guide to Good Practice* (M. Wynne, Ed.). Oxbow Books. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

Yaylali, A., Novikov, A., & Banat, H. (2021). Using corpus-based materials to teach English in K-12 settings. *The Newsletter of the Second Language Writing Interest Section*, 1–3.

Zen, E. L. (2020). A CORPUS-BASED ANALYSIS ON MULTILINGUAL CHILDREN'S NARRATIVES IN INDONESIAN CONTEXTS. *LiNGUA*, *15*(1), 91–97.

Zinsmeister, H., Heid, U., & Beck, K. (2014). Adapting a part-of-speech tagset to non-standard text: The case of STTS. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 4097–4104). European Language Resources Association (ELRA).

Zipf, G. K. (1935). *The psycho-biology of language* (1st ed.). The M.I.T. Press.

ANEXOS

Anexo I – Lista de frequência resultante da lematização do subcorpus em PE

	Lemma	Frequency
1	o	449
2	e	138
3	avião	134
4	elefantino	122
5	girafo	96
6	ele	90
7	com	89
8	um	84
9	que	80
10	a	74
11	ficar	70
12	elefante	64
13	de	57
14	brincar	50
15	muito	46
16	estar	42
17	amigo	39
18	tirar	36
19	seu	35
20	ter	34
21	para	32
22	tentar	31
23	começar	28
24	de+a	28
25	feliz	27
26	a+o	26
27	depois	26
28	mas	26
29	não	26
30	ajudar	25
31	querer	23
32	certo	23
33	triste	21
34	rede	21
35	piscina	21
36	em+a	21
37	outro	21
38	brinquedo	20
39	ir	20
40	dia	20
41	se	20
42	cair	20
43	girafito	18
44	por	18
45	deixar	18
46	chegar	18
47	conseguir	17
48	olhar	16
49	de+o	16
50	ver	16
51	vir	15
52	porque	15

	Lemma	Frequency
138	puxar	3
139	tanto	3
140	comessou	3
141	logo	2
142	mãe	2
143	bravo	2
144	observar	2
145	devolta	2
146	peixe	2
147	perceber	2
148	pérto	2
149	nada	2
150	demasiado	2
151	que-se	2
152	ella	2
153	caír	2
154	nenhum	2
155	ouvir	2
156	em+as	2
157	senhora	2
158	desistir	2
159	apanador	2
160	pechina	2
161	pedir	2
162	agúa	2
163	tava	2
164	enveja	2
165	pra	2
166	elefant	2
167	espantar	2
168	rapaz	2
169	raiva	2
170	em	2
171	chatear	2
172	estáva	2
173	assustar	2
174	consegia	2
175	admirar	2
176	algo	2
177	aproximar	2
178	fim	2
179	sempre	2
180	situação	2
181	sorrir	2
182	au	2
183	entregar	2
184	tempo	2
185	acabar	2
186	correr	2
187	preparar	2
188	homem	2
189	historiar	2

	Lemma	Frequency
275	censegoiu	1
276	roubar	1
277	outravezes	1
278	salvar	1
279	elles	1
280	segurar	1
281	seguir	1
282	apreximou-se	1
283	aprecebeu-se	1
284	ainda	1
285	passear	1
286	paz	1
287	sentir	1
288	shorar	1
289	simplesmente	1
290	emoshosa	1
291	chati-adissímo	1
292	socar	1
293	en	1
294	emquanto	1
295	perdouo	1
296	en-contrava	1
297	enclinava-se	1
298	chomou-o	1
299	cheio	1
300	aproximou-ce	1
301	tantava	1
302	encontro	1
303	picar	1
304	pichina	1
305	entre-gou-o	1
306	ciumar	1
307	aquilo	1
308	pixina	1
309	tirau	1
310	entusiasmar	1
311	claro	1
312	trousse	1
313	entusiasmo	1
314	arancou	1
315	ajudasnos	1
316	adentro	1
317	abaijouse	1
318	a+os	1
319	preguntar	1
320	problema	1
321	trubsou	1
322	errar	1
323	pé	1
324	té	1
325	esforço	1
326	comecão	1

53	chorar	14
54	contente	14
55	então	14
56	água	13
57	ser	12
58	mais	12
59	mão	12
60	encontrar	11
61	elefanta	11
62	zangar	11
63	passar	10
64	também	10
65	dar	10
66	sem	10
67	girafino	9
68	tão	9
69	isso	9
70	inveja	9
71	pescar	9
72	aparecer	9
73	á	9
74	aguar	8
75	perguntar	8
76	dizer	8
77	acontecer	7
78	girafa	7
79	giraffo	7
80	trazer	7
81	ralhar	7
82	tambem	6
83	como	6
84	pesca	6
85	poder	6
86	brincedo	6
87	até	6
88	tudo	6
89	alto	5
90	esticar	5
91	longe	5
92	pegar	5
93	enquanto	5
94	todo	5
95	elle	5
96	reparar	5
97	explicar	5
98	pensar	5
99	contar	5
100	vez	5
101	em+o	4
102	em+uma	4
103	conseguio	4
104	gritar	4
105	bem	4
106	decidir	4
190	velho	2
191	ideia	2
192	voltar	2
193	infelizmente	2
194	afastar	2
195	altura	2
196	estender	2
197	junto	2
198	ão	2
199	lado	2
200	levar	2
201	lhe	2
202	mal	1
203	de+isso	1
204	mandar	1
205	menino	1
206	de+os	1
207	de+esta	1
208	mesmo	1
209	meu	1
210	decidiou	1
211	mostrar	1
212	deishoulo	1
213	bolso	1
214	bena	1
215	ne	1
216	dejado	1
217	demora	1
218	dentro	1
219	borda	1
220	bom	1
221	obserbava	1
222	outravez	1
223	descuidadamente	1
224	braço	1
225	animação	1
226	animar	1
227	amigos	1
228	parar	1
229	pasmedos	1
230	despachar	1
231	desportivo	1
232	brincalhão	1
233	brinca-va	1
234	peiu-lhe	1
235	dexou	1
236	perdo-ou	1
237	di	1
238	apanar	1
239	aguã	1
240	perçebeu	1
241	diretamente	1
242	pesqua	1
243	dissel-lhes	1
327	quanto	1
328	veu	1
329	esperar	1
330	esperança	1
331	espremento	1
332	esplicou	1
333	arrepender	1
334	arrastar	1
335	este	1
336	começoo	1
337	álgum	1
338	éram	1
339	comigo	1
340	ashudar	1
341	éspermentar	1
342	esticavasse	1
343	esticavass	1
344	estoouse	1
345	comsigo	1
346	alcançar	1
347	alcance	1
348	etcétera	1
349	exatamente	1
350	conhecer	1
351	conceguiu	1
352	felizissimo	1
353	felissísimo	1
354	asustada	1
355	sorridente	1
356	fixe	1
357	ficarãom	1
358	sorte	1
359	só	1
360	fora	1
361	flutando	1
362	frente	1
363	consguio	1
364	atenção	1
365	alegre	1
366	adulto	1
367	terminar	1
368	gerafino	1
369	fácilemente	1
370	tirou-le	1
371	tiste	1
372	tive-se	1
373	girafa-rapaz	1
374	contentar	1
375	corar	1
376	conversar	1
377	aulto	1
378	trupesado	1
379	gostar	1
380	grande	1

107	devolver	4	244	discutir	1	381	corer	1
108	assim	4	245	dieser	1	382	grita	1
109	medo	4	246	piscino	1	383	aviar	1
110	novo	4	247	pisina	1	384	impressionar	1
111	volta	4	248	distante	1	385	cuidadósa	1
112	derepente	4	249	pois	1	386	correõ	1
113	consegiu	4	250	doio	1	387	ali	1
114	àgua	4	251	dizlhe	1	388	algum	1
115	éstava	4	252	cadinho	1	389	vieo	1
116	senhor	3	253	brinquar	1	390	vio	1
117	mulher	3	254	acondeseu	1	391	voltão	1
118	so	3	255	acompanhar	1	392	india-tamente	1
119	apanhar	3	256	pos	1	393	incontraram-se	1
120	de+as	3	257	meter	1	394	intenção	1
121	já	3	258	posivel	1	395	da-me	1
122	momento	3	259	porqué	1	396	culpar	1
123	berrar	3	260	preferir	1	397	irritar	1
124	derrepente	3	261	muit	1	398	invega	1
125	afundar	3	262	elefan-tina	1	399	aõ	1
126	pequeno	3	263	ele-fantina	1	400	isto	1
127	perto	3	264	caio	1	401	lange	1
128	buscar	3	265	pá	1	402	de+ele	1
129	quando	3	266	pâs	1	403	de+eles	1
130	ajuda	3	267	capaz	1	404	bdo	1
131	voz	3	268	elefante-menina	1	405	bambar	1
132	repente	3	269	cara	1	406	aí	1
133	estragar	3	270	aperceber	1	407	adurar	1
134	em+esse	3	271	aparesse	1	408	achar	1
135	fazer	3	272	novamente	1	409	a+a	1
136	forma	3	273	ralar	1			
137	comecou	3	274	elefantinaficou	1			

Anexo II – Concordância do lema “ficar” no subcorpus em PE

	Details	Left context	KWIC	Right context
1	doc#0	</s><s>A Elefantina viu o Girafo a brincar com avião e	**ficou**	com inveja.</s><s>Decidiu tirar-lhe o brinquedo para b
2	doc#0	dexou cair o avião.</s><s>Na piscina.</s><s>O girafo	**ficou**	triste por o seu avião que-se estava a estragar na àgua
3	doc#0	vião que-se estava a estragar na àgua.</s><s>Depois	**ficou**	zangado e ralhou com a Elefantina en voz alta.</s><s>
4	doc#0	ɛ o seu avião estava-se a afundar.</s><s>A Elefantina	**ficou**	triste por ver o seu amigo a chorar.</s><s>Depois apa
5	doc#0	› tirar o avião da àgua e deu devolta ao girafo O Girafo	**ficou**	muito contente por ter o seu brinquedo devolta e a Ele
6	doc#0	ɔntente por ter o seu brinquedo devolta e a Elefantina	**ficou**	feliz por ver o seu amigo outra vez contente.</s><s>C
7	doc#0	›ara brincar.</s><s>Depois de álgum tempo Elefantina	**ficou**	com inveja, porque tambem queria brincar com o aviã
8	doc#0	uidadósa e deixou cair o avião na água.</s><s>Girafo	**ficou**	tão zangado que começoo mandar vir com a Elefantin
9	doc#0	a fora da água, mas não o censegoiu.</s><s>O Girafo	**ficou**	muito triste e começou a chorar.</s><s>Uma elefante
10	doc#0	A elefante devolveu a o avião ao Girafo.</s><s>Girafo	**ficou**	contente por ter o avião de volta e assim tambem ficou
11	doc#0	ïcou contente por ter o avião de volta e assim tambem	**ficou**	Elefantina.</s><s>Certo dia, dois amigos incontraram-
12	doc#0	Girafo tinha um Avião de brincar.</s><s>E a Elefantina	**ficou**	com inveja e tirou o Avião de brincedo da mão do Gira
13	doc#0	ïão de brincar caiu na pixina.</s><s>e depois o Girafo	**ficou**	tanto zangado que ralhou com a Elefantina alto.</s><s
14	doc#0	ɪregou o avião de brincedo aõ Girafo.</s><s>E os dois	**ficaram**	muito felizes.</s><s>Certo dia, uma Elefante-Menina e
15	doc#0	uerer a Elefantina deixou cair o avião.</s><s>O Girafo	**ficou**	zangado e começou a ralhar com a Elefantina.</s><s>
16	doc#0	·ar o avião.</s><s>O Elefante não consegui e o Girafo	**ficou**	triste e começou a chorar Então veio uma Senhora Ele
17	doc#0	ło ao Giraffo.</s><s>O Giraffo perdo-ou a Elefantina e	**ficou**	tudo bem.</s><s>Certo dia, Girafo e Elefantina estava
18	doc#0	ina arrependeu-se "de ter ""roubado"" o avião." Giraffo	**ficou**	muito chateado com Elefantina até outro Elefante ter c
19	doc#0	a, o Girafo éstava a brincar com um avião a Elefantina	**ficou**	com ciumes e arancou o brinquedo da mão e brincou
20	doc#0	ıu com o avião.</s><s>O Girafo estava bem triste mas	**ficou**	ainda mais triste quando a Elefantina deixou o avião c
21	doc#0	ɔmeçou a berrar com a Elefantina.</s><s>A Elefantina	**ficou**	com medo mas depois vieo um outro Elefant e a Elefa
22	doc#0	· tirou o avião da agua e deu o giraffo.</s><s>O giraffo	**ficou**	bem feliz e a Elefantina bambem porque estava a ver
23	doc#0	ıois a elefantina deixou o avião cair no piscina O girafo	**ficou**	tão bravo mesmo tão bravo que começou a grita com
24	doc#0	ɬe dentro Depois ela devolveu o avião ao girafo No fim	**ficaram**	os dois felizes o girafo ficou feliz por ter o avião de volt
25	doc#0	ɔ avião ao girafo No fim ficaram os dois felizes o girafo	**ficou**	feliz por ter o avião de volta e a elefantino por o girafo
26	doc#0	·.</s><s>Corria de um lado para o outro e a Elefantina	**ficava**	a olhar.</s><s>Até que ela decidiu tirar o avião ao Gira
27	doc#0	›Até que ela decidiu tirar o avião ao Girafo.</s><s>Ele	**ficou**	com medo, porque a Elefantina o podia deixar cair.</s>
28	doc#0	estava a brincar com o seu avião.</s><s>A Elefantina	**ficou**	com inveja.</s><s>E tirou lo avião.</s><s>quando ela

29	doc#0	>sem querer deixou o avião cair a pichina.</s><s>Ela	**ficou**	triste.</s><s>A girafa gritou com a Elefantina.</s><s>[
30	doc#0	a mãe com uma pá e tirou de la o avião.</s><s>Todos	**ficarão**	feliz.</s><s>A Mãe deu a girafa o avião.</s><s>A Elef
31	doc#0	aviá brincar com o avião caiu á água.</s><s>O Girafo	**ficou**	com tanta raiva que á Elefantina tava com medo á olh
32	doc#0	o apanador de peixes.</s><s>Deu depois au Girafo e	**ficou**	muito contente.</s><s>O Girafo e á Elefantina ficarão
33	doc#0	estava a brincar com o seu avião.</s><s>A Elefantina	**ficou**	com inveja e tirou o brinquedo das mãos do Girafo e c
34	doc#0	ierer ela deixou o avião cair a piscina.</s><s>O Girafo	**ficou**	triste e zangado com a sua amiga.</s><s>Eles pedirar
35	doc#0	><s>Infelizmente, ele não conseguia.</s><s>O Girafo	**ficou**	muito triste e começou a chorar.</s><s>A Elefantina a
36	doc#0	><s>A elefanta devolveu o brinquedo ao Girafino e ele	**ficou**	novamente contente.</s><s>A Elefantina e Girafino vo
37	doc#0	vião descuidadamente deishoulo cair.</s><s>O Girafo	**ficou**	muito triste.</s><s>Como o Girafo ficou triste ele bena
38	doc#0	><s>O Girafo ficou muito triste.</s><s>Como o Girafo	**ficou**	triste ele bena aulto para a Elefantina.</s><s>Um Elef
39	doc#0	car.</s><s>Como o Gerafino não deixava a Elefantina	**ficou**	com invega e tirou o Avião das mãos da pequena Gira
40	doc#0	te o Avião caiu sem querer na Aguá.</s><s>O Girafino	**ficou**	muito zangado por a Elefantina deixar cair o Avião, e c
41	doc#0	omessou a ralhar com a voz alta com ela, a Elefantina	**ficou**	com medo.</s><s>A té que chegou um Elefante adult
42	doc#0	u no Avião e entre-gou-o o Girafino.</s><s>O Girafino	**ficou**	contente, e a Elefantina também, eles fizeram as paze
43	doc#0	rer a Elefantina deixou o avião cair na agua, O Girafito	**ficou**	tão irritado que ralhou alto a Elefantina Até que chegou
44	doc#0	neçou a chorar a Elefantina viu o seu amigo a chorar e	**ficou**	triste por ele.</s><s>Até que uma Elefante chegou cor
45	doc#0	ue o pescou (o avião) deu o Girafito.</s><s>O Girafito	**ficou**	muito feliz elles (o Girafito e a Elefantina) ficaram amig
46	doc#0	Girafito ficou muito feliz elles (o Girafito e a Elefantina)	**ficaram**	amigos outra vez e a Elefantina viu o seu amigo feliz.<
47	doc#0	nte a brincar com o seu brinquedo.</s><s>a Elefantina	**ficou**	com enveja e decidiu tirar-lhe o avião.</s><s>Infelizme
48	doc#0	a água, o que deixou o Girafo muito triste.</s><s>Ele	**ficou**	tão chateado, que comecou a berrar com ela em voz n
49	doc#0	O girafo comecou a chorar.</s><s>Já sem esperança,	**ficaram**	muito tristes.</s><s>E aí aparece mais uma elefante c
50	doc#0	nenhum... ... e deu-o ao Girafo.</s><s>Claro que ele	**ficou**	muito contente E assim o girafo e a Elefantina eram ou
51	doc#0	a ele entusiasmada.</s><s>So que então a elefantina	**ficou**	com inveja porque também queria brincar com um aviã
52	doc#0	ou caír o avião de brincar para a agúa.</s><s>Os dois	**ficaram**	a olhar muito pasmedos.</s><s>O girafo ficou tão, ma
53	doc#0	dois ficaram a olhar muito pasmedos.</s><s>O girafo	**ficou**	tão, mas tão chati-adissímo que começou a ralhar com
54	doc#0	seu alcance.</s><s>O elefante acabou por desistir e	**ficou**	sem ideias de como os podia ajudar.</s><s>O girafo fi
55	doc#0	sem ideias de como os podia ajudar.</s><s>O girafo	**ficou**	muito triste e começou a chorar.</s><s>A elefantina se
56	doc#0	ui!</s><s>Ela devolveu o avião ao.</s><s>Ele sorria e	**ficou**	muito contente.</s><s>O girafo ficou de novo com o a
57	doc#0	><s>Ele sorria e ficou muito contente.</s><s>O girafo	**ficou**	de novo com o avião.</s><s>E a girafa viu o amigos s
58	doc#0	avião.</s><s>E a girafa viu o amigos sorrir e tambem	**ficou**	muito contente.</s><s>Certo dia, um Girafo rapaz e ur
59	doc#0	fantina deixou cair o avião do amigo na piscina, Girafo	**ficou**	muito triste.</s><s>Com muita raiva, o Girafo começou
60	doc#0	<s>Quando ela o alcançou, deu o avião ao Girafo que	**ficou**	muito feliz.</s><s>O Girafo estava feliz por ter os seu
61	doc#0	ue o seu amigo estava feliz de novo, por isso também	**ficou**	muito feliz.</s><s>Certo dia, os dois amigos encontrar

62	doc#0	o O Elefante consguio e entregou o ao Girafo O Girafo	**ficou**	feliz e perdouo a Elefantina Certo dia, um Girafo e uma
63	doc#0	n avião de brincar.</s><s>Na certa altura, a Elefantina	**ficou**	com enveja, porque tambem queria brincar com o aviã
64	doc#0	Girafito.</s><s>O avião caio na agua.</s><s>O Girafo	**ficou**	triste porque pensou que o brincedo se tive-se estraga
65	doc#0	brincedo se tive-se estragado.</s><s>Depois o Girafo	**ficou**	tão zangado, que começou a gritar com a amiga.</s><
66	doc#0	o!</s><s>O Girafo estava outravez feliz e a Elefantina	**ficou**	contente de ver o seu amigo feliz.</s><s>Certo dia, a
67	doc#0	Elefantina achou o avião fixe!</s><s>Mas a Elefantina	**ficou**	com inveja e pegou o avião do Girafito.</s><s>A Elefa
68	doc#0	brincar e o avião caiu para a piscina.</s><s>O Girafito	**ficou**	zangado.</s><s>Depois de um cadinho o Girafito ficou
69	doc#0	icou zangado.</s><s>Depois de um cadinho o Girafito	**ficou**	mais zangado e ele ralou a Elefantina.</s><s>Depois
70	doc#0	e que o avião estava muito afastado.</s><s>O Girafito	**ficou**	muito triste.</s><s>Depois uma mulher veio com uma

Anexo III – Lista de frequência resultante da lematização do subcorpus em AP

	Lemma	Frequency
1	sie	311
2	die	252
3	und	124
4	giraffo	114
5	elefantina	112
6	flugzeug	111
7	eine	107
8	sein	103
9	zu	88
10	haben	69
11	mit	59
12	spielen	43
13	elefant	42
14	in	38
15	wollen	32
16	an	32
17	wasser	30
18	kommen	29
19	dann	29
20	aus	27
21	spielzeug	27
22	seine	26
23	wieder	25
24	sehr	25
25	werden	25
26	von	24
27	nicht	23
28	nehmen	23
29	aber	23
30	da	23
31	helfen	22
32	freund	21
33	dass	21
34	tag	21
35	gehen	21
36	wie	20
37	weil	20
38	was	19
39	glücklich	19
40	traurig	16
41	auch	16
42	hand	15
43	so	15
44	geben	15
45	fangen	15
46	ihre	14
47	sagen	14
48	einfach	14
49	plötzlich	13
50	fallen	12
51	froh	12
52	sehen	12

	Lemma	Frequency
147	wann	3
148	wecken	3
149	ausversehen	3
150	entschließen	3
151	jetzt	3
152	junge	3
153	kaputt	3
154	angeln	3
155	während	3
156	zwei	3
157	angst	3
158	mann	3
159	erzählen	3
160	mehr	2
161	fallte	2
162	finden	2
163	mutter	2
164	möglich	2
165	mitbringen	2
166	darüber	2
167	daraufhin	2
168	oh	2
169	nein	2
170	frech	2
171	neu	2
172	bekamm	2
173	freuen	2
174	immer	2
175	nähen	2
176	deswegen	2
177	ganz	2
178	beobachten	2
179	anschreien	2
180	reißen	2
181	rufen	2
182	gerne	2
183	geschaff	2
184	runter	2
185	geschaft	2
186	ein	2
187	gelingen	2
188	selber	2
189	schlecht	2
190	streit	2
191	giraffenjunge	2
192	geschehen	2
193	eigentlich	2
194	treffen|triefen	2
195	situation	2
196	sofort	2
197	truairg	2
198	unbedingt	2

	Lemma	Frequency
293	benschbart	1
294	bemerkt	1
295	anzuschimpfen	1
296	and	1
297	schafte	1
298	schluss	1
299	scheinen	1
300	gefallen	1
301	rausholen	1
302	schnell	1
303	geklappt	1
304	gekammt	1
305	rinnen	1
306	gelegen	1
307	schämen	1
308	ebenfalls	1
309	dort	1
310	diese	1
311	s	1
312	gerade	1
313	genommen	1
314	eifersüchtig	1
315	ecke	1
316	soffort	1
317	sobald	1
318	sollen	1
319	gesnant	1
320	geschichte	1
321	bermerkt	1
322	bereit	1
323	schon	1
324	schtrekte	1
325	schwimmbecken	1
326	soweit	1
327	gestreckt	1
328	gespanend	1
329	spielz-eug	1
330	girafe	1
331	get	1
332	einander	1
333	spilsüg	1
334	st	1
335	girafehnet	1
336	giraffino	1
337	einfersüchtig	1
338	bewundern	1
339	tauchend	1
340	traff	1
341	giraffos	1
342	girraffo	1
343	einige	1
344	somit	1

53	weit	12
54	ander	11
55	fragen	11
56	fischen	10
57	passieren	10
58	weinen	10
59	alle	10
60	schauen	10
61	auf	10
62	können	10
63	netz	9
64	neidisch	9
65	weg	9
66	probieren	9
67	versuchen	9
68	nach	8
69	böse	8
70	girafo	8
71	schimpfen	8
72	elefantin	8
73	zurück	8
74	giraffe	7
75	holen	7
76	war	7
77	wütend	7
78	erklären	7
79	beginnen	6
80	danach	6
81	schwimmbad	6
82	schaffen	6
83	schreien	6
84	strecken	6
85	um	6
86	gut	6
87	spielflugzeug	6
88	als	6
89	lassen	6
90	dan	5
91	ohne	5
92	bemerken	5
93	freunden	5
94	fröhlich	5
95	doch	5
96	einmal	5
97	bis	5
98	wassern	5
99	weiter	5
100	ob	5
101	kamm	5
102	idee	5
103	dabei	5
104	überglücklich	4
105	man	4
106	beide	4
199	beschließen	2
200	unglücklich	2
201	siht	2
202	sogar	2
203	spiel	2
204	süiel	2
205	viel	2
206	vorbei	2
207	traffen	2
208	heißen	2
209	ausverseh	2
210	bleibte	2
211	greifnetz	2
212	elefantenmädchen	2
213	badi	2
214	händ	2
215	elefand	2
216	am	2
217	w	2
218	warum	2
219	zuschauen	2
220	beckenrand	2
221	herraus	2
222	weg.	2
223	brauchen	2
224	irgent	2
225	jedoch	2
226	beeindruckend	2
227	eomes	2
228	kapput	2
229	chomt	2
230	kind	2
231	ergänzt	2
232	clever	2
233	hilf	2
234	kommte	2
235	erklerte	2
236	daher	2
237	ziehen	2
238	laufen	2
239	erschrack	2
240	leid	2
241	wegnehmen	2
242	los	2
243	lustig	2
244	erwachsen	2
245	dame	2
246	merken	1
247	etwas	1
248	mitnehmen	1
249	miteinander	1
250	essen	1
251	nahegelegen	1
252	nahe	1
345	glauben	1
346	spaß	1
347	gng	1
348	goleg	1
349	ele-fantina	1
350	einreden	1
351	bitten	1
352	beschlossen	1
353	arm	1
354	an-der	1
355	spielzeugflugzeug	1
356	spilsüig	1
357	stratt	1
358	vergessen	1
359	versinken	1
360	urhein	1
361	verzeite	1
362	groß	1
363	greifen	1
364	vorn	1
365	vorkenntnis	1
366	wait	1
367	hatt	1
368	hats	1
369	elefant-ina	1
370	elefan-tina	1
371	warin	1
372	herausnehmen	1
373	herausfischen	1
374	unter	1
375	untergehen	1
376	unvorsichtig	1
377	ur	1
378	weggenomen	1
379	wegnimmte	1
380	weggenohmen	1
381	hinzu	1
382	elefantenjung	1
383	elefanten-junge	1
384	brand	1
385	wieso	1
386	wissen	1
387	höhe	1
388	hollte	1
389	ziemlich	1
390	zeit	1
391	zunehmen	1
392	igrendwann	1
393	ig	1
394	ide	1
395	elefantinazu	1
396	übergeben	1
397	überreichen	1
398	ind	1

107	nur	4
108	frau	4
109	raus	4
110	sauer	4
111	denken	4
112	schwimmen	4
113	andere	4
114	plözlich	4
115	tun	4
116	bleiben	4
117	anfangen	4
118	flug-zeug	4
119	irgendwann	4
120	elefantiene	4
121	lehnen	4
122	laut	4
123	machen	4
124	meine	3
125	damit	3
126	namm	3
127	neben	3
128	leider	3
129	nun	3
130	netzen	3
131	nichts	3
132	denn	3
133	nochmal	3
134	passiert	3
135	für	3
136	fangnetz	3
137	schaft	3
138	pool	3
139	freundin	3
140	rennen	3
141	sieth	3
142	tauchen	3
143	elefantendame	3
144	hinaus	3
145	verspielt	3
146	vorne	3

253	natürlich	1
254	namens	1
255	fangnes	1
256	bei	1
257	nehmte	1
258	neidischund	1
259	firaffo	1
260	nett	1
261	fl-ugzeug	1
262	fischnetz	1
263	darum	1
264	nie	1
265	fliegen	1
266	noch	1
267	flugsöig	1
268	das	1
269	anschauen	1
270	anlaufen	1
271	nähe	1
272	fon	1
273	ok	1
274	oihnen	1
275	dengt	1
276	picken	1
277	freude	1
278	probiert	1
279	poll	1
280	freibad	1
281	bekommen	1
282	profiert	1
283	qint	1
284	problem	1
285	rausfischen	1
286	frsrirent	1
287	der	1
288	dentiel	1
289	richtung	1
290	frönnt	1
291	rutschen	1
292	fält	1

399	elefat	1
400	elefantine	1
401	inteligen	1
402	irgen	1
403	entfernt	1
404	bössesie	1
405	bösse	1
406	bringen	1
407	anfangte	1
408	ja	1
409	irgent-wann	1
410	entwenden	1
411	k	1
412	erfolg	1
413	entschuldigen	1
414	klein	1
415	klaver	1
416	klug	1
417	klev	1
418	erklert	1
419	erfolgen	1
420	chanstunns	1
421	landen	1
422	kraft	1
423	zurückkriegen	1
424	zusamen	1
425	legen	1
426	erschrocken	1
427	erscheinen	1
428	erstaunt	1
429	dal	1
430	lächeln	1
431	erzellte	1
432	begang	1
433	befürchten	1
434	beeindruckt	1
435	abgefal	1
436	abgefahlen	1

Anexo IV – Concordância do item “zu” no subcorpus em AP

	Details	Left context	KWIC	Right context
1	doc#0	Spielzeug dabei, ein Flugzeug.</s><s>Er fing an damit	**zu**	spielen.</s><s>Elefantina schaute ihm beeindruckt zu.<
2	doc#0	t zu spielen.</s><s>Elefantina schaute ihm beeindruckt	**zu**	.</s><s>Doch irgendwann wurde sie neidisch auf ihren
3	doc#0	gendwann wurde er böse und fing an laut mit Elefantina	**zu**	schimpfen.</s><s>Aber da kam ein anderer Elefant der
4	doc#0	schimpfen.</s><s>Aber da kam ein anderer Elefant der	**zu**	wisse schien was passiert war und wollte ihnen helfen.<
5	doc#0	siert war und wollte ihnen helfen.</s><s>Elefantina ging	**zu**	ihm und fragte ob er eine Idee hatte wie man das Flugz
6	doc#0	em Wasser holen könne.</s><s>Giraffo schaute traurig	**zum**	Flugzeug.</s><s>Der Elefant lehnte sich nach vorn am
7	doc#0	nach vorn am Beckenrand und versuchte das Flugzeug	**zu**	greifen.</s><s>Aber ohne Erfolg.</s><s>Er erklärte ihn
8	doc#0	hne Erfolg.</s><s>Er erklärte ihnen dass das Flugzeug	**zu**	weit weg schwimme.</s><s>Da fing Giraffo an zu weine
9	doc#0	ızeug zu weit weg schwimme.</s><s>Da fing Giraffo an	**zu**	weinen weil sein Flugzeug versinken würde Da kam abe
10	doc#0	ng an das Flugzeug mit dem Greifnetz aus dem Wasser	**zu**	fischen.</s><s>Als sie es aus dem Wasser gefischt hat
11	doc#0	Flugzeug spielen.</s><s>Sie entschloss es ihm einfach	**zu**	entwenden und weiter zu spielen.</s><s>Elefantina wa
12	doc#0	Sie entschloss es ihm einfach zu entwenden und weiter	**zu**	spielen.</s><s>Elefantina war so unvorsichtig, dass sie
13	doc#0	und versuchte mit aus gestreckten Armen das Flugzeug	**zu**	angeln.</s><s>Etwas dass ihm nicht gelang, da das Flu
14	doc#0	s>Etwas dass ihm nicht gelang, da das Flugzeug schon	**zu**	weit weg vom Beckenrand war.</s><s>Giraffo wurde so
15	doc#0	eckenrand war.</s><s>Giraffo wurde so traurig, dass er	**zu**	weinen anfangte.</s><s>Eine weitere Elefantin, die die
16	doc#0	sein bester freund was in den Händen hatte und siht ihn	**zu**	wie Er mit den spielzeug spielte.</s><s>Und dann wurd
17	doc#0	erte der andere Elefant das spielz-eug aus dem wasser	**zu**	holen aber er kommte nicht rann.</s><s>Und dann sag
18	doc#0	</s><s>Und dann sagte der an-dere Elefant das es viel	**zu**	weit ist.</s><s>Er kann es einfach nicht rausholen.</s>
19	doc#0	.</s><s>Und dann probierte die frau das spielzeug raus	**zu**	fischen. bis es sie irgent wann schafte.</s><s>Und gab
20	doc#0	assiert sei Der Elefant versuchte das Flugzeug herraus	**zu**	nehmen in dem er sich streckte.</s><s>Doch er schafft
21	doc#0	Deswegen entschloss sie das Flugzeug von Giraffo weg	**zu**	nehmen und selber mit das Flugzeug spielen.</s><s>G
22	doc#0	Giraffo war so traurig und böse das er anfing Elefantina	**zu**	anschreien Elefantina Erschrack sich.</s><s>Da kamm
23	doc#0	Da kamm ein anderes Elefant.</s><s>Elefantina rannte	**zum**	Elefant und sagte das sie hilfe brauchte um das Flugzeu
24	doc#0	und sagte das sie hilfe brauchte um das Flugzeug weg	**zu**	nehmen vom Wasser.</s><s>Sie dachten alle darüber r
25	doc#0	dann sahen sie das der Elefant versuchte den Flugzeug	**zu**	fischen aber es ging nicht weil es zu weit war daher sag
26	doc#0	chte den Flugzeug zu fischen aber es ging nicht weil es	**zu**	weit war daher sagte er: "Es geht leider nicht" da kam e
27	doc#0	s><s>Giraffo blieb so wütend und fang an mit Elefantina	**zu**	schimpfen Danach kam ein anderen Elefant und Elefant
28	doc#0	en Danach kam ein anderen Elefant und Elefantina ging	**zu**	ihn zu.</s><s>Das Elefant wollte die 2 freunde helfen. d

29	doc#0	ach kam ein anderen Elefant und Elefantina ging zu ihn	**zu**	.</s><s>Das Elefant wollte die 2 freunde helfen. dann le
30	doc#0	ın lehnte er sich nach vorne und probierte das Flugzeug	**zu**	holen.</s><s>Aber er hat es nicht geschaff Giraffo bleib
31	doc#0	inen netz Mit dem Netz probierte sie das Flugzeug raus	**zu**	angeln. sie hat es geschaft und gab das Flugzeug an G
32	doc#0	Deswegen entschloss sie das Flugzeug von Giraffo weg	**zu**	nehmen und selber mit das Flugzeug spielen.</s><s>G
33	doc#0	Giraffo war so traurig und böse das er anfing Elefantina	**zu**	anschreien Elefantina Erschrack sich.</s><s>Da kamm
34	doc#0	Da kamm ein anderes Elefant.</s><s>Elefantina rannte	**zum**	Elefant und sagte das sie hilfe brauchte um das Flugzeu
35	doc#0	t und sagte das sie hilfe brauchte um das Flugzeug weg	**zu**	nehmen vom Wasser.</s><s>Sie dachten alle darüber r
36	doc#0	dann sahen sie das der Elefant versuchte den Flugzeug	**zu**	fischen aber es ging nicht weil es zu weit war daher sag
37	doc#0	chte den Flugzeug zu fischen aber es ging nicht weil es	**zu**	weit war daher sagte er: "Es geht leider nicht" da kam e
38	doc#0	></s><s>Giraffo blieb so wütend und fang an mit Elefantina	**zu**	schimpfen Danach kam ein anderen Elefant und Elefant
39	doc#0	en Danach kam ein anderen Elefant und Elefantina ging	**zu**	ihn zu.</s><s>Das Elefant wollte die 2 freunde helfen. d
40	doc#0	ach kam ein anderen Elefant und Elefantina ging zu ihn	**zu**	.</s><s>Das Elefant wollte die 2 freunde helfen. dann le
41	doc#0	ın lehnte er sich nach vorne und probierte das Flugzeug	**zu**	holen.</s><s>Aber er hat es nicht geschaff Giraffo bleib
42	doc#0	inen netz Mit dem Netz probierte sie das Flugzeug raus	**zu**	angeln. sie hat es geschaft und gab das Flugzeug an G
43	doc#0	isch.</s><s>Sie nehmt den Flugzeug weck und fang an	**zu**	spielen.</s><s>Sie lässt dem Flugzeug in das Wasser f
44	doc#0	ihn die Situation.</s><s>Er hilft, aber das Flugzeug war	**zu**	weit weck und er Kann das nicht von das Wasser weck
45	doc#0	reck nehmen.</s><s>Giraffo war unglücklich, er beginnt	**zu**	weinen.</s><s>Elefantina war schämen.</s><s>Plötzlic
46	doc#0	st.</s><s>Elefantina sagt passieren Der Elefant probiert	**zum**	helfen. aber er könnte nicht Giraffo weint.</s><s>Dann
47	doc#0	Elefant sie isst inteligen sie hatte ein Sie probiert mit den	**zum**	den flugzeug fangen.</s><s>Sie hatte die flugzeug und
48	doc#0	n ist, sie schauten ihn an.</s><s>Elefantina ging soffort	**zu**	ihn und erzellte was passiert ist.</s><s>Das Elefant stre
49	doc#0	nicht fangen.</s><s>Girafo war sehr traurig und fing an	**zu**	weinen, Elefantina blieb sehr traurig, dass er weint.</s>
50	doc#0	gab den Spielzeug Girafo, er war sehr glücklich.</s><s>	**Zum**	schluss waren Girafo und Elefantina wieder freunde.</s
51	doc#0	wollte ihr Fl-ugzeug haben, aber dann ist der Flugzeug	**zur**	pool abgefalen.</s><s>Giraffo war sehr traurig, das der
52	doc#0	eug) sehr wait ist und, dass er schaft nicht der Flugzeug	**zur**	nehmen.</s><s>Elefantina ist traurig dass Giraffo weint
53	doc#0	zlich eine Elefat.</s><s>Sie probiert auch der Flugzeug	**zu**	nehmen.</s><s>Sie schaft es und gibt der Flugzeug an
54	doc#0	wachsener Elefant vorbei.</s><s>Elefantina kam sofort	**zu**	ihm gelaufen.</s><s>Sie erklärte ihm was passiert sei u
55	doc#0	hnen helfen könne.</s><s>Der Elefant beschloss ihnen	**zu**	helfen.</s><s>Er probierte sich so weit zu strecken wie
56	doc#0	schloss ihnen zu helfen.</s><s>Er probierte sich so weit	**zu**	strecken wie möglich, jedoch ohne erfolg Der Elefant er
57	doc#0	ne erfolg Der Elefant erklärte ihnen, dass das Flugzeug	**zu**	weit entfernt schwimme.</s><s>Und da, wie aus dem n
58	doc#0	al was geschehen war und die Dame beschloss oihnen	**zu**	helfen.</s><s>Sie kam wieder mit einem Fischnetz ang
59	doc#0	e kam wieder mit einem Fischnetz angelaufen um ihnen	**zu**	helfen.</s><s>Und es hatte geklappt Sie übergab Giraf
60	doc#0	</s><s>Er wollte ihnen helfen.</s><s>Elefantina rannte	**zu**	ihm sie fragte ihm ob er eine Idee hätte, wie man das Fl
61	doc#0	ı sollte.</s><s>Er wollte sich strecken um das Flugzeug	**zu**	holen.</s><s>Er streckte sich soweit wie möglich.</s><

62	doc#0 aber nicht.</s><s>Der Giraffo wurde traurig und begann	**zu**	weinen.</s><s>Die Elefantina sah wie sie ihren Freund
63	doc#0 lücklich gemacht hat.</s><s>Eine kluge Elefantine kam	**zu**	ihren.</s><s>Sie hatte einen Fangnetz mitgebracht, um
64	doc#0 ugzeug dabei.</s><s>Girafo begann mit den Flug-zeug	**zu**	spielen.</s><s>Elefantina schaute nur zu wie seinen fre
65	doc#0 len Flug-zeug zu spielen.</s><s>Elefantina schaute nur	**zu**	wie seinen freund mit den Flugzeug spielte.</s><s>Plöz
66	doc#0 >Giraffo war plözlich so böse und begann mit Elefantina	**zu**	schimpfen, sie war erstaunt.</s><s>Plötzlich bekamm E
67	doc#0 i freunden helfen.</s><s>Elefantina ging schnell auf ihn	**zu**	und fragte ob er eine idee hätte, den Flugzeug aus den
68	doc#0 gte ob er eine idee hätte, den Flugzeug aus den Wasser	**zu**	nehmen.</s><s>Der Elefant begann an zu helfen, er sc
69	doc#0 den Wasser zu nehmen.</s><s>Der Elefant begann an	**zu**	helfen, er schtrekte sich, aber es ging nicht.</s><s>Der
70	doc#0)er Elefant dann, erklärte die Kinder, dass der spielzeug	**zu**	weit weg war.</s><s>Aber plözlich tauchte eine klavere
71	doc#0 nder helfen würde.</s><s>Sie begann an der Flug-zeug	**zu**	fischen, die anderen schauten nur zu, wie sie es machte
72	doc#0 an der Flug-zeug zu fischen, die anderen schauten nur	**zu**	, wie sie es machte.</s><s>Als sie den Flugzeug hatte,
73	doc#0 /s><s>Plötzlich fing Giraffo an mit seinem Spielflugzeug	**zu**	spielen.</s><s>Elefantina begang einfersüchtig zu werd
74	doc#0 :eug zu spielen.</s><s>Elefantina begang einfersüchtig	**zu**	werden, weil sie auch mit dem Flugzeug spielen wollte.
75	doc#0 pielen wollte.</s><s>Da nahm Elefantiene Giraffo ohne	**zu**	bitten einfach das Spielflugzeug aus der Hand fing an m
76	doc#0 einfach das Spielflugzeug aus der Hand fing an mit dem	**zu**	spielen.</s><s>Auf einmal fiel Elefantiene das Flugzeug
77	doc#0 rsuchte das Flugzeug aus dem Wasser mit seiner Hand	**zu**	holen.</s><s>Giraffo fing an zu weinen, weil er sich ein
78	doc#0 Wasser mit seiner Hand zu holen.</s><s>Giraffo fing an	**zu**	weinen, weil er sich einredete er würde sein Flugzeug n
79	doc#0 /s><s>Sie schaute dem Giraffen Jungen beeindruckend	**zu**	, wie er damit spielte Irgendwann wurde Elefantina neid
80	doc#0 ar und wollte ihnen helfen.</s><s>Elefantina lief auf ihn	**zu**	.</s><s>Sie fragte ob er eine Idee hätte, wie man den S
81	doc#0 wie sein Flugzeug unterging.</s><s>Die Freunde sahen	**zu**	, wie sich der Elefant nach vorne lehnte um das Flugzeu
82	doc#0 nt nach vorne lehnte um das Flugzeug aus dem Wasser	**zu**	fischen.</s><s>Aber er schaffte es nicht, da der Flugze
83	doc#0 chen.</s><s>Aber er schaffte es nicht, da der Flugzeug	**zu**	weit weg war.</s><s>Da fing Giraffo an zu weinen.</s>
84	doc#0 ler Flugzeug zu weit weg war.</s><s>Da fing Giraffo an	**zu**	weinen.</s><s>Da merkte Elefantina, dass sie ihren Fre
85	doc#0 evere Elefantin auf.</s><s>Sie hatte beschlossen ihnen	**zu**	helfen und hatte einen Netz in der Hand.</s><s>Währe
86	doc#0 anschauten, fing sie an das Flugzeug aus dem Wasser	**zu**	fischen.</s><s>Sobald sie es hatte gab sie das Spielze
87	doc#0 eine Frau mit einem "Netz".</s><s>Sie war bereit Ihnen	**zu**	helfen.</s><s>Während Sie es versuchte raus zu picke
88	doc#0 Ihnen zu helfen.</s><s>Während Sie es versuchte raus	**zu**	picken waren alle 3 sehr froh!</s><s>Ja!</s><s>Die Fra