TÉCNICO LISBOA

PE2LGP 4.0: de português europeu para língua gestual portuguesa

Matilde do Carmo Lages Gonçalves

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur
Prof. Hugo Miguel Aleixo Albuquerque Nicolau

Júri

Presidente: Prof. Francisco António Chaves Saraiva de Melo

Orientador: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur

Vogal: Prof. Ricardo Daniel Santos Faro Marques Ribeiro

Setembro 2020

Agradecimentos

Esta tese não teria sido concretizada com a mesma qualidade sem o conhecimento partilhado, os momentos passados e o apoio recebido durante a sua realização.

Obrigada aos meus orientadores, Luísa Coheur e Hugo Nicolau, pelo apoio, orientação, incentivo e por me socorrerem nos momentos de maior aflição, pelas reuniões descontraídas e divertidas que desfizeram o medo e o nervosismo inicial de fazer uma tese e pela confiança em mim.

Obrigada ao grupo de investigação do projeto *Corpus Linguístico e avatar da LGP (PTDC/LLT-LIN/29887/2017)* do Instituto de Ciências da Saúde da Universidade Católica Portuguesa pela oportunidade de integrá-lo como linguista computacional, pela receção terna, pelo incentivo e suporte que me deram desde o primeiro dia, pela disponibilidade em ajudar-me e em colaborar e pelo conhecimento partilhado sobre linguística, língua gestual portuguesa e cultura dos surdos, que contribuíram para a compreensão das características e fenómenos da língua gestual portuguesa e para a escrita técnica da tese.

Obrigada amigos e colegas, especialmente Carol e Pedro pelo apoio e pelos divertidos momentos que certamente não serão esquecidos.

Obrigada namorado por me amparares sempre que caio, me incentivares, me ajudares e me acalmares todos os dias.

Obrigada família por teres acreditado em mim.

Obrigada Sky por seres a minha bola anti-stress.

Muito obrigada!

Abstract

Like all natural languages, Portuguese Sign Language evolved naturally, acquiring grammatical characteristics different from Portuguese. Therefore, the development of a translator between the two languages consists in more than a mapping of words into signs (which results in a form of signed Portuguese), as it should ensure that the translation it produces satisfies the grammar of Portuguese Sign Language. Previous works use only manual translation rules and are very limited in the amount of grammatical phenomena that they cover, producing signed Portuguese. This thesis presents the first translation system from Portuguese to Portuguese Sign Language based not only on manual rules, but also on translation rules automatically built from grammatical information annotated in a corpus, the reference corpus under development by Universidade Católica Portuguesa. The manual rules deal with grammatical phenomena that the translation rules do not cover, namely morphological phenomena, such as the marking of the female gender and integrate particularities of the language such as facial expressions. It is the first work that deals with grammatical facial expressions that mark interrogative and negative sentences. Given a sentence in Portuguese, the system returns a sequence of glosses with markers that identify facial expressions, spelled words, among others. The thesis reports both a manual and an automatic evaluation. Results show improvements in the quality of the translation compared to the baseline system based on signed Portuguese.

Keywords

Portuguese Sign Language; Automatic translation; Annotated corpus; Data-driven machine translation; Rule-based translation system; Glosses; Natural language processing

Resumo

A língua gestual portuguesa, tal como a língua portuguesa, evoluiu de forma natural, adquirindo características gramaticais distintas do português. Assim, o desenvolvimento de um tradutor entre as duas não consiste somente no mapeamento de uma palavra num gesto (português gestuado), mas em garantir que os gestos resultantes satisfazem a gramática da língua gestual portuguesa. Trabalhos desenvolvidos anteriormente utilizam exclusivamente regras de tradução manuais, sendo muito limitados na quantidade de fenómenos gramaticais abrangidos, produzindo pouco mais que português gestuado. Nesta dissertação desenvolveu-se o primeiro sistema de tradução de português para língua gestual portuguesa, que para além de regras manuais, se baseia em regras de tradução construídas automaticamente a partir de informações gramaticais anotadas num corpus, o corpus de referência em desenvolvimento pela Universidade Católica Portuguesa. As regras manuais tratam de fenómenos gramaticais que as regras de tradução não cobrem, nomeadamente fenómenos morfológicos, como a marcação do género feminino e integram particularidades da língua como as expressões faciais. É o primeiro trabalho que lida com as expressões faciais gramaticais que marcam as frases interrogativas e negativas. Dada uma frase em português, o sistema devolve uma sequência de glosas com marcadores que identificam expressões faciais, palavras soletradas, entre outros. Uma avaliação automática e uma avaliação manual são apresentadas, indicando os resultados melhorias na qualidade da tradução em comparação ao sistema *baseline* (português gestuado).

Palavras Chave

Língua gestual portuguesa; Tradução automática; Corpus anotado; Sistema de tradução baseado em regras; Sistema de tradução baseado em corpus; Glosa; Processamento da língua natural;

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Acrónimos

NLP	processamento de língua natural
LGP	língua gestual portuguesa
LP	língua portuguesa
PE	português europeu
SVO	sujeito-verbo-objeto
TER	translation error rate
BLEU	bilingual evaluation understudy
NER	reconhecimento de entidades nomeadas
AM	análise morfossintática
SOV	sujeito-objeto-verbo
SVO	sujeito-verbo-objeto
SV	sujeito-verbo
FCT	Fundação para a Ciência e a Tecnologia

1

Introdução

Conteúdo

As línguas gestuais são conhecidas como línguas visuo-espaciais, i.e., a comunicação é realizada por gestos produzidos em determinados locais no espaço tridimensional ou no corpo. Os gestos são constituídos por uma componente manual e não manual (expressões faciais e movimentos corporais).

A língua gestual portuguesa (LGP) é a principal forma de comunicação entre a comunidade surda portuguesa. Um tradutor de português para LGP pode ser usado para facilitar a comunicação entre ouvintes e surdos, e também para fins de aprendizagem da LGP. No entanto, por ser uma língua natural, possui diferenças gramaticais em relação à língua portuguesa (LP), como na ordem das palavras e na estrutura frásica base [1]. Assim, para que um tradutor não produza apenas "português gestuado" (tradução em que cada palavra em português é directamente transformada num gesto em LGP, sem obedecer às suas regras gramaticais) terá de ter em conta as características gramaticais da LGP. Apesar de existirem alguns estudos linguísticos sobre esta, não existe ainda uma gramática oficial, nem sequer consenso sobre variados fenómenos linguísticos. Por exemplo, sobre a estrutura frásica base, alguns autores consideram que é sujeito-verbo-objeto (SVO), outros sujeito-objeto-verbo (SOV). Talvez por isso os poucos trabalhos computacionais ligados à tradução para LGP [2–6] focam pouco a componente linguística, baseando-se em pequenos conjuntos de regras manuais e excluindo expressões faciais, resultando em pouco mais do que português gestuado. De modo a colmatar estas falhas e a impulsionar a criação de recursos computacionais para o processamento automático da LGP, o projecto "Corpus & Avatar da Língua Gestual Portuguesa"[1], liderado pela Universidade Católica Portuguesa, está a criar o primeiro corpus linguístico de referência da LGP. Neste, as unidades lexicais são transcritas em glosas e anotadas com informações gramaticais. Com esta dissertação contribui-se com um tradutor para LGP, em que a(s) frase(s) traduzida(s) para LGP são representadas por sequências de glosas, com marcadores que identificam as expressões faciais e palavras soletradas. O sistema de tradução apoia-se em regras de tradução automáticas e num dicionário bilingue criados automaticamente a partir do corpus de referência. Na base da tradução encontra-se ainda um conjunto de regras manuais que capturam fenómenos linguísticos relacionados com a morfologia das palavras, como a marcação do feminino, e integram particularidades que as regras automáticas não cobrem, tais como as expressões faciais. Segundo o levantamento do trabalho relacionado realizado, este é o primeiro tradutor para LGP com uma forte componente linguística e que, em particular, lida com expressões faciais gramaticais essenciais para marcar frases interrogativas e negativas.

Duas avaliações foram realizadas, uma automática com base num corpus de teste construído por um especialista e outra manual, em que falantes de LGP avaliam a qualidade das traduções.

[1] PTDC/LLT-LIN/29887/2017

1.1 Objetivos

Os principais objetivos deste trabalho são:

- Desenvolver uma componente de extração de informações gramaticais do corpus de referência e, a partir dessas informações, construir regras de tradução e um dicionário bilingue.
- Criar uma componente de tradução de frases em português europeu (PE) para LGP tendo por base as regras de tradução e o dicionário bilingue extraídos do corpus e regras manuais.
- Estabelecer uma notação de gestos em glosa para marcar informações adicionais sobre expressões faciais, palavras soletradas, entre outras, baseada nas convenções usadas no corpus.
- Avaliar diferentes ferramentas de processamento da língua natural para texto em português tais como: etiquetadores morfossintáticos e reconhecedores de entidades com o intuito de discriminar as ferramentas com melhor desempenho para serem usadas na componente de tradução.

1.2 Contribuições

Os recursos desenvolvidos nesta dissertação serão tornados públicos. As principais contribuições deste trabalho são:

1. Desenvolvimento do primeiro sistema de tradução de texto em português para LGP baseado em regras de tradução e um dicionário bilingue construídos automaticamente a partir do corpus de referência (podendo crescer com o corpus);
2. Um conjunto de regras manuais;
3. Informações com valor linguístico, como:
    - Regras de tradução automáticas, que descrevem as ordens dos constituintes morfossintáticos e frásicos na LGP em relação à LP.
    - Estatísticas sobre as regras automáticas, nomeadamente a frequência de cada uma no corpus de referência.
4. Avaliação em grande escala de ferramentas *open-source* nas diferentes tarefas de processamento de língua natural (NLP) para texto em português, com o objetivo de facilitar a escolha da ferramenta que melhor se adequa à tarefa em mãos.
5. Recursos usados na avaliação das ferramentas de NLP: corpora de teste para cada tarefa de NLP, *scripts* de avaliação e de conversão entre as etiquetas dos corpora de teste e as etiquetas de cada ferramenta.

6. O sistema é adequado para servir de *baseline* para trabalhos futuros por ter um corpus de teste e uma avaliação automatizada que permite avaliar rapidamente modificações ao sistema.

1.3 Artigos produzidos

Com o conhecimento adquirido ao longo desta dissertação vários artigos foram produzidos:

1. *PE2LGP: traduzindo português europeu para língua gestual portuguesa* (Matilde Gonçalves, Luísa Coheur, Hugo Nicolau e Ana Mineiro): artigo em produção para a revista Linguamática.
2. *Avaliação de recursos computacionais para o Português* (Matilde Gonçalves, Luísa Coheur, Jorge Baptista e Ana Mineiro): artigo submetido na revista Linguamática.
3. *Entre o gesto e a glosa: Critérios de categorização de classes de gestos de um corpus de referência da Língua Gestual Portuguesa* (Mara Moita, Matilde Gonçalves, Helena Carmo, Sebastião Palha, Neide Gonçalves, Paulo Carvalho, Celda Morgado e Ana Mineiro): comunicação oral e resumo submetido na conferência *III Encontro sobre Morfossintaxe da LGP e de outras línguas de sinais*.
4. *PE2LGP Animator: A Tool To Animate A Portuguese Sign Language Avatar* (Pedro Cabral, Matilde Gonçalves, Ruben dos Santos, Hugo Nicolau e Luisa Coheur): artigo publicado em *9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*.
5. *The Opposite Signs in Portuguese Sign Language: a phono and morphological analysis* (Sebastião Palha, Matilde Gonçalves e Mara Moita): artigo submetido na conferência *Formal and Experimental Advances in Sign Language Theory*.

1.4 Estrutura do documento

Este documento está organizado em mais seis capítulos: no **Capítulo 2** são apontados alguns aspetos da gramática da língua gestual portuguesa e é apresentado o corpus de referência. A revisão da literatura encontra-se no **Capítulo 3**. No **Capítulo 4** avaliam-se ferramentas *open-source* de processamento de texto em português. No **Capítulo 5**, o sistema de tradução desenvolvido é descrito e no **Capítulo 6** apresentam-se a metodologia de avaliação e a análise dos resultados. No último capítulo (**Capítulo 7**) são feitas considerações finais.

2

Background

Conteúdo

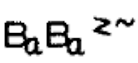
Nas secções seguintes são apresentados os principais sistemas de escrita da língua gestual (Secção 2.1), as diferenças gramaticais entre LGP e a LP (Secção 2.2) e o corpus em desenvolvimento pela Universidade Católica (Secção 2.3).

2.1 Sistemas de anotação de gestos

Uma das exigências na tradução de línguas orais para línguas gestuais e na construção de um corpus bilingue é o uso de uma notação que represente os gestos. De seguida, são descritos de forma breve, os principais sistemas de anotação:

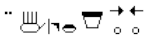
- **Glosa** – Os gestos são anotados usando as palavras com o mesmo significado na língua falada, mas em letras maiúsculas [7]. Por exemplo, a palavra *coelho*, em LGP, será anotada com a glosa *COELHO*. Este sistema não requer conhecimento de símbolos para descrever os gestos e informações sobre gestos não manuais, como expressões faciais, poderão ser adicionadas.
- **Sistema de notação de Stokoe** – Esta notação foi criada por William Stokoe, em 1960, para a língua gestual americana. É um sistema de transcrição fonético composto por letras que descrevem a configuração da mão e por símbolos que transcrevem o movimento e localização da mão. Uma limitação do sistema de Stokoe é a exclusão dos gestos não manuais na notação [8].
- **SignWriting** - SignWriting foi desenvolvida em 1974. As componentes dos gestos manuais (localização, configuração, movimento do gesto e orientação) e os elementos não manuais são representados usando símbolos icónicos [9].
- **Sistema de notação de Hamburgo (HamNoSys)** - É um sistema de transcrição fonético, em que a notação de um gesto consiste na descrição simbólica dos seus fonemas: componentes manuais (configuração, movimento e orientação das mãos e posição do gesto) e não manuais (com a nova versão). Ao contrário das notações anteriores, HamNoSys permite representar gestos produzidos com as duas mãos [10].

Na Figura 2.1 encontra-se um exemplo de um gesto transcrito nas diferentes notações[1].



(a) Na notação de Stokoe.

(b) Na notação de SignWriting.



(c) Na notação de Hamnosys.

Figura 2.1: Representação do pronome interrogativo *Qual* na língua gestual americana.

[1] Este exemplo foi retirado de www.signwriting.org/forums/linguistics/ling001.html

2.2 Gramática da língua gestual portuguesa

Os primeiros estudos sobre a LGP surgiram na década de 90, pelo que pouco se sabe sobre a língua. Por não existir uma gramática oficial, não há consenso sobre vários aspetos gramaticais, nomeadamente sobre a ordem base ou ordem canónica[2] das frases.

2.2.1 Ordem frásica base

As frases na língua portuguesa seguem a estrutura sintática básica SVO, mas em LGP a ordem é diferente. Alguns autores defendem que a estrutura predominante é SOV [11]. No entanto, em 2016, realizou-se um estudo sobre a ordem básica das frases declarativas que pretendeu responder às seguintes questões [1]: a) A LGP possui alguma ordem básica dos constituintes? Se possuir, qual é? e b) Quais são os fatores que influenciam a ordem dos constituintes? Para responder a estas perguntas foram feitas duas experiências, com surdos e falantes. Concluiu-se que existe uma ordem sintática básica e que é igual à da língua portuguesa (SVO). Contudo, e respondendo à segunda questão, podem ser construídas frases com outras ordens mas, para que sejam perceptíveis, têm que ser acompanhadas por gestos não manuais. A topicalização[3] e frases de pergunta-resposta são alguns aspetos referidos no estudo que justificam o uso de outras ordens de palavras. Este estudo foi realizado apenas com verbos transitivos não locativos e para frases declarativas, pelo que esta informação será comparada e completada com dados retirados do corpus descrito na Secção 2.3.

2.2.2 Tipos de frases

O tipo de frase, se é interrogativa ou negativa, influencia a ordem dos seus constituintes. De acordo com [1], as frases interrogativas são marcadas pelo uso de advérbios e pronomes interrogativos, no final de uma frase LGP, acompanhadas sempre pela expressão facial interrogativa.

2.2.3 Género feminino

A marcação do género feminino nos nomes em LGP é realizada pela composição de gestos, ou seja, pela adição do gesto que marca o género, o gesto *MULHER*, ao gesto base. O gesto sem marcação de género está, por omissão, no género masculino [1]. Assim, o gesto *LEÃO* como é um substantivo masculino é representado apenas pelo gesto *LEÃO* enquanto que *LEOA* é composto por *MULHER* + *LEÃO*. É importante notar a ordem de produção dos gestos: primeiro, tem-se o gesto marcador de género feminino e depois o gesto principal. Contudo, existem situações em que não há marcação do

[2] A ordem canónica estabelece-se a partir de frases declarativas e afirmativas, que não tenham sofrido topicalização.

[3] Topicalização acontece quando uma parte da frase é colocada no início de modo a ser destacada. Por exemplo, dada a frase *Eu tenho medo desses cães*, o objeto (*desses cães*) pode ser destacado na frase, ficando *Desses cães, eu tenho medo*.

género em nomes que se referem a animais por existirem gestos para cada nome: existem gestos diferentes para os nomes *galo* e *galinha*, por exemplo [12]. Também existem situações em que os gestos no feminino e masculino têm a mesma configuração e orientação da mão, mas a sua produção é realizada em locais diferentes do corpo, por exemplo, os gestos *ENFERMEIRO* (produzido no ombro) e *ENFERMEIRA* (produzido na testa).

2.2.4 Diminutivo e aumentativo

À semelhança da marcação do género feminino, a representação do diminutivo e aumentativo é feita pela composição de gestos, mais precisamente com a adição dos gestos *PEQUENO* e *GRANDE*, respetivamente, ao gesto base. A produção desses gestos tem que ser acompanhada de expressões faciais para marcar o grau do substantivo. Assim, *LEOAZINHA* é composto pelos gestos *MULHER* + *LEÃO* + *PEQUENO* (com expressão facial). Os gestos que marcam o grau do substantivo são produzidos por último.

2.2.5 Plural

Existem quatro formas para marcar o plural [1]:

(i) repetição do gesto usando a mão dominante. Por exemplo, para o gesto *ÁRVORES* repete-se o gesto *ÁRVORE*;

(ii) adição de um numeral que normalmente procede o substantivo. Por exemplo, *cinco livros* corresponde à sequência dos gestos *LIVRO* + *CINCO*;

(iii) adição de um advérbio de quantidade. Por exemplo, *muitos livros* corresponde a *LIVRO* + *MUITO*.

(iv) redobro, o gesto é produzido e repetido com as duas mãos. Um exemplo comum, é o caso do plural de *pessoa*, com as duas mãos repete-se o gesto *PESSOA* simultaneamente.

2.2.6 Determinantes possessivos

Em LGP, determinantes possessivos (*meu*, *teu*, etc.) procedem o substantivo [1, 4]. Por exemplo, a frase *o teu irmão* originará a sequência de gestos: *IRMÃO* + *TEU*.

2.2.7 Determinantes artigos, verbos copulativos e nomes próprios

Os determinantes artigos definidos e indefinidos e os verbos *ser* e *estar* não são representados em LGP. Os nomes próprios são soletrados, caso não tenha sido atribuído um nome gestual prévio à entidade referida pelo nome.

2.2.8 Preposições

As preposições não são representadas em LGP isoladamente [13], algumas são incorporadas no movimento dos gestos para identificar, por exemplo, os locais inicial e final do objeto que está em movimento [1].

2.2.9 Conjunções coordenadas

De acordo com o estudo preliminar sobre conexões interfrásicas e frásicas [14], as conjunções coordenadas adversativas (*mas* e *porém*) são lexicais, ou seja são produzidas manualmente, enquanto que a conjunção coordenativa copulativa *e* é uma conexão prosódica, expressa não manualmente. A expressão predominante associada a esta conjunção é a expressão facial neutra.

2.2.10 Tempos verbais

Em LGP, os verbos são aplicados no modo infinitivo. Segue-se informação sobre a sua marcação, retirada do livro "Um olhar sobre a morfologia dos gestos" [12].

A marcação do presente pode ser realizada de duas maneiras: recorre-se apenas ao modo infinitivo do verbo ou acrescenta-se advérbios de tempo ao modo infinitivo do verbo.

A marcação dos tempos verbais passado e futuro realiza-se de três formas [12]: pela adição de expressões faciais à forma neutra do verbo (modo infinitivo do verbo); pela adição de advérbios de tempo (ontem, amanhã, etc.) no início da frase, caso estes existam na frase, caso contrário, adicionam-se no início da frase os gestos *PASSADO* ou *FUTURO*.

2.2.11 Negação

De acordo com os autores do artigo [15] existem dois tipos de negação em LGP, a negação regular e a negação irregular. Na primeira, a marcação da negação é realizada pela adição de marcadores gramaticais de negação manuais como por exemplo, a adição do gesto manual *NÃO* ou do gesto *NADA* depois do verbo ou pela adição de gestos não manuais, como o marcador de negação *headshake* (abanar a cabeça de um lado para o outro repetidamente) ou a alteração da expressão facial. Na negação irregular, a negação está incorporada no verbo, i.e., existem gestos diferentes para a negação de um certo verbo (por exemplo, *NÃO-QUERER* e *QUERER*).

2.2.12 Classificadores

De acordo com [16], os classificadores são unidades gestuais que possuem uma estrutura semântico-sintática complexa. Existem duas categorias de classificadores: os nominais e os verbais. Os primeiros

especificam características de um referente (objeto ou pessoa), como informações aspetuais e locativas. Por exemplo, existe um gesto classificador nominal para *pessoa* associado a uma determinada configuração da mão. Os segundos, incorporam ações nesses referentes. Por exemplo, os gestos para *pintar com rolo* e *pintar com lápis*, são produzidos de forma diferente. Uma descrição mais detalhada sobre os classificadores encontra-se no artigo [16]. O artigo mencionado é um estudo piloto sobre classificadores. Por ser uma estrutura complexa, a sua geração automática foi deixada para trabalho futuro.

2.3 Corpus anotado para LGP

Na Universidade Católica Portuguesa, está a ser desenvolvido um corpus anotado para língua gestual portuguesa, por um grupo constituído por seis membros, 2 linguistas com conhecimento em LGP, três especialistas em LGP (2 deles surdos) e uma intérprete. O corpus é constituído por vídeos de surdos portugueses de diferentes faixas etárias (dos 10 aos 60 anos), contendo discursos formais, não formais, espontâneos ou com assunto previamente estabelecido. As anotações são realizadas com o *software* ELAN[4], uma ferramenta que permite a criação de várias camadas de anotações de vídeos e áudio. Estas camadas estão alinhadas e sincronizadas temporalmente com o vídeo ou áudio. Neste corpus, estão a ser anotados a tradução da mensagem enunciada no vídeo para língua portuguesa, os gestos transcritos em glosa, as respetivas classes gramaticais e os argumentos da oração (sujeito e objeto). Na anotação com glosa são seguidas convenções para identificar as informações gramaticais (Tabela 2.1) e os diferentes fenómenos linguísticos (Tabela 2.2).

Classe gramatical	**Convenção**
Substantivo	N
Verbo	V
Adjetivo	ADJ
Advérbio	ADV
Elemento sintático	**Convenção**
Argumento externo	ARG_EXT
Argumento interno	ARG_INT

Tabela 2.1: Exemplos de convenções usadas na anotação de informações gramaticais.

Fenómeno gramatical	**Convenção (exemplo)**
Datilologia	**DT(**M**-**A**-**R**-**I**-**A**)**
Gesto composto[5]	POR**-**FAVOR
Flexão em género	**FG(**MULHER**+**GATO**)**

Tabela 2.2: Exemplos de convenções usadas na anotação de fenómenos linguísticos.

[4] tla.mpi.nl/tools/tla-tools/elan

3

Trabalho Relacionado

Conteúdo

Nesta secção, apresenta-se a investigação dos principais tópicos para o projeto. Começa-se por detalhar dois tipos de formalismos gramaticais. De seguida. são enunciados os diferentes tipos de tradução automática e particulariza-se os tradutores automáticos desenvolvidos para LGP. Por último, enumeram-se ferramentas para a componente de tradução destacando as suas funcionalidades.

3.1 Gramáticas síncronas

Nesta secção, descrevem-se algumas gramáticas síncronas usadas na tradução automática, nomeadamente, na transferência gramatical entre um par de línguas. As gramáticas síncronas apresentadas são formalismos gramaticais, i.e., um conjunto de regras que descrevem como uma frase pode ser produzida seguindo a gramática da língua. Estas gramáticas síncronas consistem na definição simultânea das regras gramaticais da língua origem e da língua alvo. Descrevem como uma frase com uma determinada sintaxe e léxico pode ser transformada numa frase na língua alvo, respeitando a sua gramática.

3.1.1 Formalismos gramaticais

3.1.1.A Gramáticas livres de contexto síncronas

As *Synchronous context-free grammars* (SCFG) ou, em português, gramáticas livres de contexto síncronas, foram originalmente introduzidas por Lewis e Stearns para a compilação de linguagens de programação (conhecidas como *syntax-directed transduction grammars*) [17]. Este formalismo é uma extensão às gramáticas livres de contexto (CFGs) na medida em que as SCFGs especificam regras de sintaxe que pertencem a gramáticas livres de contexto, para duas línguas simultaneamente. De modo a facilitar a compreensão das SCFGs serão explicadas, primeiro, as gramáticas livres de contexto.

As gramáticas livres de contexto são um tipo de formalismo gramatical constituído por um conjunto de regras que descrevem as categorias sintáticas de uma língua e o léxico [18]. Informalmente, as gramáticas livres de contexto são definidas por um conjunto: de símbolos não terminais, que são abstrações dos símbolos terminais; de símbolos terminais que correspondem às palavras presentes na linguagem ou às etiquetas morfossintáticas; de regras na forma: $A \rightarrow \beta$, onde A pertence ao conjunto dos símbolos não terminais e β uma sequência de um ou mais símbolos não terminais ou terminais. S é o símbolo inicial da gramática e pertence ao conjunto dos símbolos não terminais.

Exemplo 3.1.1. Considere-se o seguinte conjunto de regras:

(1) $S \rightarrow SN \;\; SV$

(2) $SN \rightarrow DET \;\; NOM$

(3) $SV \rightarrow VI$

(4) $DET \rightarrow a|o$

(5) $NOM \rightarrow Rita|Jorge$

(6) $VI \rightarrow adormeceu$

1) A primeira regra define que uma frase é constituída por um sintagma nominal (SN) seguido de um sintagma verbal (SV), ou seja, S pode ser reescrito por $SN\ \ SV$.

2) Por sua vez, um SN é constituído por um determinante (DET) seguido de um nome (NOM), o que quer dizer que SN pode ser reescrito por $DET\ \ NOM$.

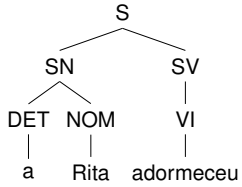
3) Um SV é composto apenas por um verbo intransitivo.

4) A partir desta regra são especificados os símbolos terminais. Nesta linguagem um DET poderá ser reescrito por *a* ou por *o*.

5) Da mesma forma, um nome poderá ser reescrito por *Rita* ou por *Jorge*.

6) Um verbo intransitivo será reescrito por *adormeceu*.

No lado esquerdo das regras, i.e., à esquerda da seta ($\rightarrow$), encontra-se, sempre, um único símbolo não terminal e no lado direito uma sequência de um ou mais símbolos terminais ou não terminais. A aplicação sequencial das regras é conhecida como derivação. A forma mais comum de representar uma derivação é usando árvores de análise sintática. De seguida, encontra-se a árvore de análise sintática da derivação completa da frase "a Rita adormeceu", do exemplo anterior.



Existem vários algoritmos que atribuem uma estrutura sintática a uma frase de acordo com as regras da gramática livre de contexto. É o caso do algoritmo de Cocke-Kasami-Younger (CKY) e do algoritmo de Earley (**Secção 3.1.2**).

Nas gramáticas livres de contexto síncronas a derivação é feita de forma simultânea entre as duas regras livres de contexto. No exemplo 3.1.2, listam-se as regras síncronas necessárias para a transferência gramatical da frase em (7), que está em língua portuguesa, para a ordem sintática correta em LGP, originando a sequência de glosas em (8). Para este exemplo, assume-se que a ordem correta de uma frase em LGP é SOV.

Exemplo 3.1.2.

(7) O João come a sopa

(8) JOÃO SOPA COMER

(9) S $\rightarrow$ <SN2 SV3, SN2 SV3>

(10) SN $\rightarrow$ <DET7 NOM8, DET7 NOM8>

(11) SV $\rightarrow$ <V5 SN6, SN6 V5>

(12) DET $\rightarrow$ < o | a, ϵ >

(13) NOM $\rightarrow$ <João | sopa, JOÃO | SOPA>

(14) V $\rightarrow$ <come, COMER>

Em contraste com as regras CFG, as regras livres de contexto síncronas possuem dois lados direitos separados por uma vírgula: o lado origem e o lado alvo, que correspondem respetivamente à língua da qual queremos traduzir e à língua para a qual queremos traduzir. Além disso, as regras SCFGs são assinaladas com um número natural, que permite marcar a correspondência bijetiva entre os símbolos não terminais do lado origem com os símbolos não terminais do lado alvo das regras síncronas. Só é possível derivar regras síncronas que tenham símbolos não terminais ligados [19]. Por exemplo, o símbolo $SN2$ do lado origem está ligado ao $SN2$ do lado alvo, tornando possível a derivação simultânea desses dois símbolos não terminais. À semelhança das CFGs, a derivação das SCFGs pode ser representada por árvores de análise sintática. A diferença é que serão originadas duas árvores de análise sintática porque são feitas duas derivações, uma no lado origem e a outra no lado alvo. As árvores de análise sintática do exemplo 3.1.2 estão representadas na Figura 3.1.

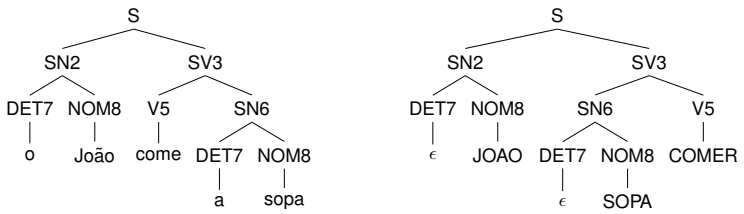


Figura 3.1: Árvores de análise sintática que resultam da derivação síncrona das regras do exemplo 3.1.2.

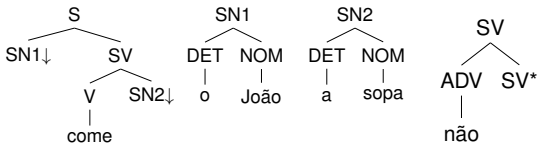
As árvores de análise sintática partilham a mesma estrutura mas diferem na ordem dos símbolos não terminais.

Uma particularidade das SCFGs é que permitem a reordenação, apenas, entre nós irmãos, i.e., nós que estejam no mesmo nível hierárquico sob o mesmo nó pai [20]. Se fosse do interesse derivar uma frase com a estrutura OSV, já não seria possível a troca entre sujeito (SN2) e objeto (SN6). Uma solução para este problema é tornar a árvore menos profunda e mais plana ou usar outro formalismo que resolva essa limitação como as *Synchronous tree-adjoining grammars*.

3.1.1.B *Synchronous Tree-Adjoining grammars*

Este formalismo é uma variante das Tree-Adjoining grammars (TAGs) que foram propostas por Joshi em 1975 e estendidas para a área de língua natural por Vijay-Shankar e Joshi, em 1985 [21].

À semelhança dos outros formalismos, as TAGs são definidas por um conjunto de símbolos terminais, não terminais, um símbolo inicial e por dois tipos de árvores elementares: as árvores iniciais e as árvores auxiliares, que representam as regras da gramática. As árvores iniciais e auxiliares são formadas por: nós interiores[1], que são nós não terminais, e por nós folhas[2], que podem ser nós terminais ou nós não terminais. A diferença entre as duas é marcada pela identificação dos nós folhas não terminais, enquanto que, nas árvores iniciais estes são identificados por uma seta direcionada para baixo ($\downarrow$), nas árvores auxiliares identificam-se com um asterisco e são chamados de *nós de anexo*. Na Figura 3.2 encontra-se um exemplo de árvores elementares que compõem as TAGs.



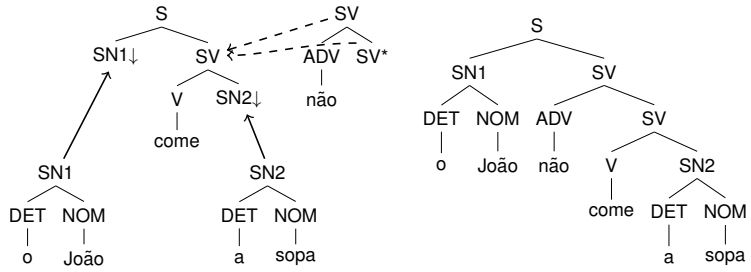
(a) Árvores iniciais. **(b)** Árvore auxiliar.

Figura 3.2: Exemplo de árvores elementares das TAGs

Essas árvores elementares são combinadas através de duas operações: a operação de substituição e a operação de adjunção. Na operação de substituição, um nó folha não terminal de uma árvore inicial é substituído por uma árvore elementar com o mesmo símbolo não terminal na raiz que o nó não terminal a substituir. A operação de adjunção consiste em anexar uma árvore auxiliar numa árvore inicial em qualquer nó não terminal (seja folha ou não folha) [22]. A derivação nas TAGs inicia-se com as árvores iniciais às quais são aplicadas, a cada passo, uma das operações ou a de substituição ou a de adjunção.

Pegando nas árvores do exemplo anterior, estas podem ser combinadas através das duas operações apresentadas anteriormente. Os nós folhas não terminais ($SN1$ e $SN2$), representados por uma seta, da árvore inicial podem ser substituídos pelas árvores elementares que representam esses sintagmas nominais, i.e., que têm como raiz o mesmo símbolo que os nós não terminais ($SN1$ e $SN2$), por via da operação de substituição. Noutro passo, recorrendo à operação de adjunção, ao nó não terminal SV da árvore inicial é anexada a árvore auxiliar que representa o advérbio de negação (*não*). Este exemplo de derivação e o resultado da mesma estão representados na Figura 3.3.

[1] Nós interiores correspondem aos nós que não são nós folhas, i.e., que possuem nós filhos.
[2] Nós folhas de uma árvore são nós que não possuem nós filhos.



(a) As setas a tracejado representam as operações de adjunção, enquanto que, as outras representam substituições.

(b) Resultado da derivação.

Figura 3.3: Exemplo de derivação nas TAGs aplicando as operações de substituição e de adjunção à árvore inicial.

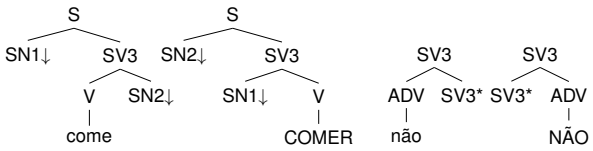
No caso das STAGs são definidas duas TAGs para as línguas origem e alvo. Assim, cada árvore elementar nas TAGs corresponde a um par de árvores elementar nas STAGs. A derivação nas STAGs realiza-se com os mesmos princípios que nas TAGs, mas serão produzidas duas árvores de forma síncrona, em vez de uma. Ou seja, as operações de substituição e de adjunção são aplicadas simultaneamente ao par de árvores elementares. Mais uma vez é importante que os nós não terminais das duas árvores estejam ligados entre si para a derivação ser válida [23]. Assim, para que a frase *O João não come a sopa* seja traduzida para a seguinte frase em LGP *SOPA JOÃO COMER NÃO* são necessários os pares de árvores elementares representados na Figura 3.4, para a língua portuguesa e para a língua gestual portuguesa. O resultado da derivação encontra-se na Figura 3.5.

A operação de substituição deste formalismo estende o domínio de localidade em relação às SCFGs, permitindo a reordenação entre nós que não sejam irmãos. No caso do exemplo na Figura 3.1, com as STAGs é possível a troca entre sujeito e objeto. Por sua vez, a operação de adjunção torna possível a criação de uma frase mais complexa a partir de uma estrutura base e anexando outras árvores elementares a essa estrutura, por exemplo, de uma frase afirmativa gera-se a frase negativa.

Tal como no formalismo anterior, a transferência gramatical realiza-se através da análise sintática da frase da língua origem aplicando as regras da gramática. Nas SCFGs os algoritmos analisam uma frase com base em regras livres de contexto síncronas mas, nas STAGs, as regras são representadas por árvores, fazendo emergir outras versões dos algoritmos CKY e Earley. Na secção seguinte, estes algoritmos são brevemente descritos.

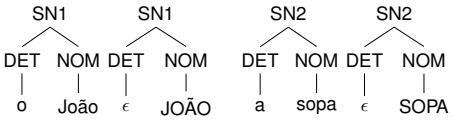
3.1.2 Algoritmos de análise sintática

Dada uma frase numa língua, a respetiva frase na língua alvo pode ser gerada usando um dos algoritmos que se seguem. Estes algoritmos, recorrendo a estratégias diferentes, aplicam as regras da



(a) Par de árvores iniciais.

(b) Par de árvores auxiliares.



(c) Par de árvores elementares para a operação de substituição.

Figura 3.4: Exemplo de árvores elementares nas STAGs

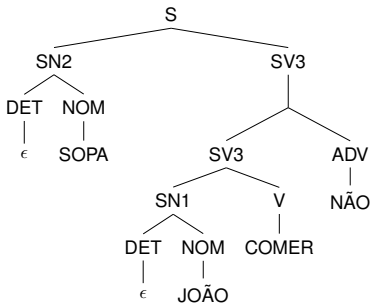


Figura 3.5: Resultado da derivação das árvores da figura anterior usando as operações de adjunção e de substituição.

gramática síncronas à frase de origem e simultaneamente são derivadas duas árvores sintáticas que representam a frase na língua origem e a sua tradução.

3.1.2.A Algoritmo CKY

O algoritmo CKY usa uma abordagem de programação dinâmica[3] e uma estratégia *bottom-up* para executar a análise sintática da frase. No entanto possui uma restrição quanto às regras de derivação, estas deverão estar na forma normal de Chomsky, ou seja, no lado direito de cada regra deverão existir dois símbolos não terminais ou apenas um símbolo terminal.

Forma normal de Chomsky [18]: $A \rightarrow BC$, $A \rightarrow \alpha$, em que A, B e C são símbolos não terminais e α é um símbolo terminal. Nas situações em que não é possível a conversão para a forma normal de Chomsky, o CKY não pode ser usado. No entanto, existem outros algoritmos nos quais não é imposta essa restrição, como é o caso do algoritmo de Earley. Mais informações sobre a implementação do

[3]Programação dinâmica é uma técnica algorítmica que permite guardar os resultados de um subproblema evitando que estes sejam recalculados.

algoritmo para as CFGs podem ser encontradas em [18].

Vijay-Shanker e Joshi, em 1985, estenderam este algoritmo para as TAGs [24]. Este algoritmo limita o número de ramos das árvores elementares das TAGs para dois.

3.1.2.B Algoritmo Earley

O algoritmo Earley segue uma abordagem de programação dinâmica e uma estratégia *top-down* que permite realizar a análise sintática de uma frase que pertence a uma gramática livre de contexto. Este algoritmo não exige que as regras estejam na forma normal de Chomsky, ao contrário do CKY. Detalhes da sua implementação são descritos em [25].

Em 1988, Yves Schabes e Aravind K. Joshi desenvolveram uma versão do algoritmo Earley adaptado às TAGs que segue as estratégias *bottom-up* e *top-down*. A sua implementação encontra-se descrita nos seguintes artigos [26] e [24].

3.1.3 Sistema XTAG

O sistema XTAG[4] é um *workbench* gráfico para o desenvolvimento de TAGs implementado na linguagem *Common Lisp*. Fornece uma interface com ferramentas que permitem a definição de gramáticas, a edição das árvores elementares, a derivação das árvores com o algoritmo de análise sintática, entre outras funcionalidades. O projeto possui também uma componente para a transferência gramatical usando as STAGs, o que permite a tradução entre um par de línguas. Atualmente, este projeto inclui a implementação das gramáticas para o coreano e para o inglês, cobrindo uma grande parte dos fenómenos linguísticos [27].

3.2 Tradução automática

A área de tradução de língua natural apresenta inúmeros desafios devido à variabilidade linguística e à ambiguidade (lexical, sintática, etc.) entre línguas naturais. Várias técnicas têm sido desenvolvidas para contornar esses desafios e automatizar o processo de tradução. Realçam-se os métodos baseados em regras manuais e os baseados em *corpora*. Nas duas últimas décadas, os sistemas baseados em corpora têm retirado a popularidade aos primeiros métodos devido à disponibilidade de extensos corpora e pela qualidade superior das traduções [28].

[4] Site oficial do projeto XTAG: www.cis.upenn.edu/ xtag/

3.2.1 Tradução automática baseada em regras

Os primeiros tradutores automáticos comercializados implementavam um sistema de tradução baseada em regras manuais com informações morfológicas, sintáticas e semânticas do par de línguas [29]. Neste tipo de tradução, linguistas especificam um conjunto de regras que permitem a transferência gramatical entre a língua origem e a língua alvo. Sistemas que implementem este tipo de tradução, conseguem captar vários fenómenos linguísticos, no entanto, implicam um grande esforço por parte dos linguistas na criação das regras, que é dificultada com a presença de exceções linguísticas [29].

Vários sistemas de tradução automática baseados em regras para língua gestual foram propostos, tais como ATLASLang [30], tradução de discurso espanhol para língua gestual espanhola [31], tradução para língua gestual ucraniana em telemóveis [32] e VLibras [33]. A maioria usa a abordagem baseada na transferência gramatical (transferência sintática, lexical e semântica) através de regras de tradução criadas por linguistas.

Em TEAM [34], um protótipo de um sistema de tradução de texto inglês para língua gestual americana, as regras de tradução são definidas usando *tree-adjoining grammars* (Secção 3.1.1.B), resolvendo divergências linguísticas como a ordem das palavras nas frases. As regras síncronas são representadas em pares de árvores elementares. Esses pares de árvores expressam tanto a componente manual como a não manual de um gesto. Neste protótipo, a frase de entrada possui informações gramaticais como o tipo de frases e informações morfológicas como o tempo verbal, que são incorporadas na derivação da árvore da língua alvo por denotarem gestos não manuais necessários para uma tradução correta. O resultado da componente de transferência gramatical corresponde a uma sequência de glosas com parâmetros embebidos, como expressões faciais, que serão passados a um avatar para a produção dos gestos. Não foi encontrada uma avaliação deste sistema.

Em [35] é proposto um sistema baseado em regras para traduzir texto árabe para língua gestual árabe (não detalham o dialeto estudado). Este trabalho descreve uma notação baseada em glosas para representar os gestos, realçando a importância de uma notação para a transcrição de gestos num tradutor. No sistema de notação baseado em glosa adotado são adicionados marcadores para discriminar, entre outros, palavras que deverão ser soletradas, identificadas através do símbolo # no início e no fim da palavra, como em #MARIA#. A frase de entrada em árabe é analisada morfológica, sintática e semanticamente e, com base nessas informações, são aplicadas regras manuais definidas de acordo com as diferenças gramaticais entre as duas línguas. O resultado do sistema é uma sequência de glosas que respeita a estrutura e gramática da língua gestual árabe e a sequência respetiva de imagens dos gestos. A avaliação do sistema, usando um corpus paralelo com frases relacionadas com o domínio da saúde, revelou uma acurácia de 82%.

O tradutor automático descrito em [36] apresenta uma solução para os classificadores na língua gestual espanhola. O artigo descreve um sistema com uma arquitetura da abordagem baseada na

transferência gramatical que traduz texto em espanhol para uma sequência de glosas, respeitando a gramática e estrutura da língua gestual espanhola. Ao texto em espanhol é feita uma análise sintática dos constituintes e de dependências, resultando numa árvore de dependências. Dada essa árvore, inicia-se a fase de transferência do léxico e da estrutura, usando respetivamente um dicionário bilingue e regras que permitem transformar a árvore de dependências da frase espanhola para a correspondente árvore de dependências da língua gestual espanhola. É nesta fase que os classificadores nominais são identificados recorrendo às relações semânticas entre as palavras (hiperonímia e holonímia). Por exemplo, as árvores de fruto, na língua gestual espanhola, como *laranjeira* são compostas pelo classificador nominal *árvore* seguido do nome do fruto, neste caso *laranja*. Por fim, na fase da geração são aplicadas regras morfológicas e da ordem das palavras à árvore de dependências, resultando numa sequência de glosas que representa a tradução na língua gestual espanhola. Nesta fase, os classificadores verbais são identificados. Na língua gestual espanhola, os classificadores verbais correspondem a sintagmas nominais com modificadores preposicionais e locuções prepositivas com informações locativas (*sobre* e *atrás*) e temporais (*entre* e *desde*).

3.2.2 Tradução baseada em dados de corpora

Neste tipo de tradução inserem-se a tradução automática estatística e baseada em redes neuronais[5]. Para treinar um modelo estatístico é usado um extenso *corpus* bilingue. O desempenho dos sistemas de tradução baseados em algoritmos de aprendizagem depende da qualidade, tamanho e domínio dos dados do corpus. Apesar das dificuldades em criar um corpus para línguas gestuais, alguns sistemas de tradução foram desenvolvidos com base em modelos estatísticos. É o caso destes sistemas de tradução para a língua gestual americana [37] e alemã [38].

Destaca-se o trabalho desenvolvido por Hung-Yu Su e Chung-Hsien Wu [39]. Os autores apresentam um sistema de tradução estatístico de texto em mandarim para língua gestual de taiwan (TSL) que lida com a escassez de dados num corpus paralelo. A transferência gramatical baseia-se num formalismo gramatical, mais precisamente, em regras síncronas de gramática livre de contexto e numa memória de tradução que descreve a ordem dos papéis temáticos entre as frases de ambas as línguas. A estrutura sintática das frases em TSL e a memória de tradução são extraídas do corpus bilingue através do alinhamento entre o léxico das frases do corpus bilingue. As palavras e os gestos são alinhados usando uma medida de semelhança, em vez de métodos probabilísticos. Para a avaliação, foram traduzidas 25 frases curtas (menos de 10 palavras) e 25 frases compridas retiradas de livros escolares chineses por dois sistemas: o presente sistema e um sistema de tradução estatística desenvolvido por Chiu e outros [40]. Os resultados mostram que o procedimento exposto supera o segundo sistema, usando o mesmo corpus pequeno, principalmente para frases extensas. Este sistema não identifica

[5] As redes neuronais são modelos computacionais usados na aprendizagem automática.

a componente não manual dos gestos. A estratégia implementada para o alinhamento de palavras e gestos neste tradutor foi uma fonte de inspiração para o presente sistema, dado que a gramática é igualmente extraída de um corpus de pequenas dimensões, a partir do qual o treino de um modelo de alinhamento não seria possível.

3.3 Sistemas de tradução automática para LGP

Quanto à LGP, alguns tradutores automáticos foram desenvolvidos, todos baseados num conjunto de regras manuais.

3.3.1 PE2LGP: de português europeu para língua gestual portuguesa

A primeira versão deste projeto, foi concretizada por Inês Almeida em 2014 [2, 41]. O sistema de tradução de texto para LGP é constituído por três componentes: processamento da língua natural (análise morfossintática, reconhecimento de entidades e análise de dependência) e tradução em glosa; mapeamento entre a glosa produzida no módulo anterior e as animações correspondentes e, por último, produção de gestos usando um avatar 3D implementado no Blender[6]. Quanto ao módulo de tradução para glosa, este segue uma abordagem baseada em regras, em que são transferidos o léxico e a estrutura LGP. Na transferência lexical a palavra poderá ser transformada em glosa ou soletrada no caso de ser uma entidade. Na transferência sintática, os constituintes da frase são reordenados com a aplicação de regras de tradução. Por falta de informações que permitissem construir regras sobre o plural, este não foi explorado nesta versão. A avaliação foi realizada com falantes e surdos de duas associações portuguesas de surdos. Consistiu na interpretação de 3 palavras e uma frase animadas pelo avatar 3D.

A segunda versão (PE2LGP 2.0) [3] visa agilizar o processo de criação de gestos para a sua produção num avatar.

PE2LGP 3.0 [5] é a terceira versão deste projeto, que se foca na componente de tradução, mais precisamente, no processamento da língua natural e na sua transformação numa representação intermédia de texto (SIGML) que contém as informações necessárias para animar um avatar. O sistema de tradução é composto por dois módulos: *Sign tokenization* e *SIGML Generation*. No primeiro módulo, são extraídas informações morfossintáticas e feita uma análise de dependência sintática com as quais será realizada a transferência gramatical para LGP, juntamente com um corpus paralelo. O corpus paralelo contém para cada estrutura das frases em português, a respetiva estrutura sintática em LGP com classificadores e expressões faciais anotados. A transferência gramatical consiste, então, na corres-

[6]Blender é um programa que permite, entre outras tarefas, a modelação e animação de modelos tridimensionais. Mais informações disponíveis em www.blender.org

pondência total da estrutura sintática da frase de entrada com uma das estruturas sintáticas em língua portuguesa presente no corpus paralelo, usando expressões regulares. Se houver correspondência, a estrutura da frase de entrada será substituída pela estrutura em LGP respetiva. Caso contrário, a frase resultante terá a mesma estrutura que a frase de entrada em português. O problema é que só foram anotadas 63 regras, o que não abrange muitos fenómenos da LGP. Assim, a probabilidade da estrutura de uma frase de entrada ser correspondida por uma existente no corpus é muito baixa, resultando em frases traduzidas para português gestuado e não para LGP, visto que a sintaxe da LGP não foi cumprida. Esta limitação implica que frases negativas e tempos verbais não sejam corretamente traduzidos e que classificadores e expressões faciais sejam excluídos da tradução. Como output deste módulo, tem-se um ficheiro de *Sign tokens* na ordem válida para LGP e com informações morfológicas. Esse ficheiro será a entrada para o segundo módulo (SIGML Generation). Para a geração do ficheiro SIGML, cada token é primeiro associado ao seu ficheiro SIGML, recorrendo a um dicionário de ficheiros SIGML e a um conjunto de regras. Estas regras permitem resolver os casos de palavras que são soletradas, de classificadores e de algumas palavras no género feminino. O ficheiro SIGML final serve para alimentar o avatar JaSigning[7], que produzirá os gestos. Na avaliação foram avaliados três aspetos: o número de frases dadas como entrada que tiveram uma correspondência com as estruturas no corpus paralelo, a facilidade de inserir novos gestos no dicionário SiGML e a percetibilidade e precisão dos gestos produzidos pelo avatar. Concluiu-se com a avaliação do módulo de transferência gramatical que algumas frases tinham etiquetas de dependência sintática incorretas atribuídas pela ferramenta de NLP (Spacy) e que devido ao número reduzido de regras, a probabilidade de uma frase aleatória ser corretamente traduzida é baixa. O sistema, também, não é capaz de traduzir substantivos com flexão de grau e número.

3.3.2 IF2LGP: intérprete automático de fala em língua portuguesa para língua gestual portuguesa

Este sistema [4] permite a tradução de fala para língua gestual portuguesa, usando vídeos retirados do YouTube. A mensagem produzida oralmente é traduzida usando uma abordagem baseada em regras. A componente de tradução está dividida em dois módulos: no primeiro módulo é realizada a análise morfossintática da frase e, no segundo, estão definidas as regras da LGP. Para a identificação das regras no segundo módulo, foram usadas apenas 10 frases. Dada uma frase de entrada, esta é pré-processada no primeiro módulo, onde são identificadas as etiquetas morfossintáticas e no segundo módulo são aplicadas as regras de tradução, de acordo com as etiquetas morfossintáticas e com a posição da palavra na frase, por exemplo, se a palavra for um numeral então seguirá o substantivo.

[7] JaSigning é um avatar virtual que produz os gestos com base num ficheiro SIGML. Encontram-se mais informações no site: vh.cmp.uea.ac.uk/index.php/JASigning.

Com estas duas informações, é efetuada a reordenação dos constituintes e são tratadas palavras no feminino e no plural, de acordo com a gramática da LGP. Depois da aplicação das regras, uma sequência de palavras em glosa é retornada e, com recurso a uma base de dados de vídeos, cada gesto é associado ao respetivo vídeo. Assim, o resultado do sistema de tradução corresponde a uma sequência de vídeos com a animação de cada palavra. Foi usado o software FFMPEG para concatenar os vídeos automaticamente. A avaliação da conversão para uma sequência de glosas foi realizada por intérpretes e revelou resultados positivos. Com vídeo é difícil integrar as expressões faciais, fazendo com que algumas frases deixassem de ser inteligíveis. Apesar dos resultados positivos, constata-se que este sistema não identifica particularidades da LGP como as expressões faciais e classificadores e não está preparado para traduzir substantivos no diminutivo ou aumentativo.

3.3.3 VIRTUALSIGN

VIRTUALSIGN[8] [6] é um sistema de tradução bidirecional de texto, realizado com o objetivo de simplificar a aprendizagem e a comunicação entre surdos e não surdos. Os autores não detalham a abordagem seguida na tradução de texto português para língua gestual portuguesa, mas a transferência lexical é realizada diretamente, ou seja, fazendo corresponder cada palavra da frase a um gesto guardado numa base de dados com informações sobre a sua animação necessária para alimentar o avatar 3D. Não se encontrou informações sobre a avaliação deste projeto.

3.4 Ferramentas para o processamento da língua portuguesa

Neste trabalho, o critério de escolha para o conjunto de possíveis ferramentas para o processamento de texto em português baseia-se em três características, têm que ser: *open source*, compatíveis com Python e válidas para textos em português europeu. As 8 ferramentas encontradas que cumprem este critério são: Freeling, Natural language toolkit (NLTK), NLPyPort, OpenNLP, Polyglot, Spacy, StanfordNLP e TreeTagger. No capítulo 4, as características destas ferramentas são descritas em detalhe e é realizada uma avaliação minuciosa do seu desempenho nas diferentes tarefas de NLP.

[8] Página do projeto VIRTUALSIGN: 193.136.60.223/virtualsign/pt/index.php

4

Avaliação de Ferramentas de NLP

Conteúdo

Neste capítulo, as ferramentas *open-source* listadas anteriormente (Secção 3.4) serão avaliadas em relação a duas tarefas de NLP: análise morfossintática (AM) e reconhecimento de entidades nomeadas (NER). As ferramentas que se destacarem nesta avaliação serão usadas na componente de processamento de texto português do sistema de tradução proposto.

Nos dias de hoje, a área do NLP encontra-se em profunda expansão e em Portugal não é excepção. Actualmente várias empresas têm projetos neste campo, desenvolvendo sistemas de pesquisa em dados médicos, agentes virtuais, sistemas de tradução, etc. Do mesmo modo, vários agentes interessados utilizam ferramentas que operam sobre o português e, se a língua inglesa continua a ser imbatível em termos de recursos disponíveis, existem actualmente várias ferramentas gratuitas que oferecem modelos pré-treinados (ou facilmente treináveis) para a língua portuguesa, em especial para as variantes do português europeu e do Brasil. Coloca-se, então, a questão de escolher a ferramenta mais adequada para a tarefa em mãos. Não é do conhecimento, a existência na literatura de uma avaliação em grande escala destas ferramentas nas diferentes tarefas de NLP para texto em português, que permita que essa questão seja respondida. Tal resultará da complexidade do processo de avaliação destes sistemas, que vai desde da instalação de cada ferramenta até à conversão das etiquetas, passando pela escolha ou anotação de um conjunto de dados para teste (criação da coleção dourada). Por esta razão decidiu-se avaliar várias ferramentas disponíveis publica e gratuitamente para a língua portuguesa em duas tarefas de NLP, de forma a que não-especialistas decidam rapidamente que ferramentas usar consoante a sua finalidade.

Das informações conseguidas nesta avaliação escreveu-se e submeteu-se um artigo para a revista Linguamática[1] (ainda em revisão). O objetivo não é comparar detalhadamente as várias ferramentas e escolher a vencedora com base em sofisticados corpora de referência anotados, mas ter uma ideia da utilidade de uma ferramenta e/ou modelos associados, com base numa metodologia correta, consoante as necessidades da aplicação final em vista. Assim, além dos valores de desempenho de cada ferramenta nas diferentes tarefas de NLP, mostra-se a facilidade de instalação e de utilização de cada ferramenta. Além das informações expostas neste capítulo, no artigo submetido apresenta-se, ainda uma avaliação qualitativa, realizada por um linguista de duas ferramentas (SpaCy e Stanford) que realizam a tarefa de análise de dependências, tendo em conta vários factores, tais como a pontuação, a segmentação e a etiquetagem morfossintática, dos quais dependem as dependências obtidas.

As contribuições desta avaliação e do artigo são: a construção de dois *corpora* de referência, um para cada uma das tarefas (AM e NER); a adaptação dos *corpora* de referência tendo em conta os diferentes pré-processamentos dos dados, realizados pelas diversas ferramentas; os *scripts* de conversão entre as etiquetas de cada ferramenta e as etiquetas dos *corpora* de referência; a avaliação de sete ferramentas (onze modelos diferentes) na tarefa de análise morfossintática; a avaliação de sete

[1] https://linguamatica.com/index.php/linguamatica

modelos distintos na tarefa de NER e a avaliação (qualitativa) de dois analisadores na tarefa de análise de dependências.

4.1 Ferramentas e recursos em análise

Nesta secção, são descritas as ferramentas e modelos pré-treinados, desenvolvidos para o processamento de português, para as tarefas de análise morfossintática e NER e disponibilizados gratuitamente. A maioria destas ferramentas fornece modelos pré-treinados para as tarefas em estudo. As linguagens de programação destas ferramentas alternam entre o Java, o C++ e o Python, e, de um modo geral, apresentam documentação, o que torna relativamente fácil a sua instalação e utilização.

4.1.1 Natural Language Toolkit - NLTK

NLTK [42] é uma plataforma para a linguagem de programação Python que facilita o acesso a *corpora* e oferece várias bibliotecas para diferentes tarefas de NLP, como divisão em *tokens*, análise morfossintática, NER e papéis semânticos. O uso desta plataforma requer algum conhecimento em programação. Contudo, existe um conjunto de programas para a linha de comandos chamado NLTK-Trainer[2] que abstrai o utilizador da programação, facilitando o treino de modelos presentes na ferramenta NLTK, a avaliação desses modelos e a análise de *corpus*.

A instalação da plataforma NLTK encontra-se documentada para cada sistema operativo[3]. NLTK não fornece modelos pré-treinados, no entanto, disponibiliza *corpora* anotados para serem treinados. Os corpora disponíveis para o processamento de texto em português fazem parte do projeto "Floresta Sintática"[4]. A utilização das componentes que a ferramenta oferece é facilitada com a publicação de alguns exemplos[5] para as diferentes tarefas de processamento de texto em português. Os corpora Floresta Sintática não contêm informação sobre entidades nomeadas, pelo que a realização da tarefa de NER depende de outros corpora para texto em português ou da criação de novos, caso seja do interesse treinar um modelo ou, ainda, recorrer-se a modelos pré-treinados.

4.1.2 NLPyPort

NLPyPort[6] [43] é uma ferramenta direcionada para o processamento de texto português e baseada em modelos e funções da ferramenta NLTK. Realiza NER, análise morfossintática, identificação dos

[2]NLTK-Trainer é acessível em https://github.com/japerk/nltk-trainer e documentado em https://nltk-trainer.readthedocs.io/en/latest/

[3]A sua documentação encontra-se em https://www.nltk.org/install.html

[4]Esses corpora estão disponíveis em https://www.linguateca.pt/Floresta/.

[5]Estes exemplos encontram-se em http://www.nltk.org/howto/portuguese_en.html

[6]https://github.com/jdportugal/NLPyPort

lemas e disponibiliza uma versão melhorada da função de divisão em tokens da ferramenta NLTK, na qual os clíticos e contrações são devidamente tratados. De notar que existe uma versão anterior desta ferramenta compatível com projetos desenvolvidos na linguagem de programação Java [44].

4.1.3 Polyglot

Para o processamento de texto em português destacam-se as seguintes tarefas: divisão de tokens, NER e a análise morfossintática. Para cada uma, a ferramenta Polyglot [45] disponibiliza modelos pré-treinados, inclusive para português. Por ser uma biblioteca para Python, a sua utilização pressupõe experiência com programação. A documentação[7] desta ferramenta encontra-se bem estruturada e inclui tutoriais e informações sobre cada tarefa, além dos procedimentos para a sua instalação e importação.

4.1.4 SpaCy

SpaCy[8] [46] é uma biblioteca para o NLP que incorpora modelos estatísticos pré-treinados de várias línguas, inclusive de português. As tarefas disponíveis correspondem à classificação morfossintática e análise de dependências. Esta ferramenta foi pensada para ser importada como biblioteca em programas Python e não disponibiliza outras opções de uso. A sua documentação contém, entre outras informações, procedimentos para a sua instalação, importação e realização das diferentes tarefas de NLP acompanhados por exemplos. No entanto, não existe documentação sobre as etiquetas usadas na análise morfossintática.

4.1.5 Freeling

FreeLing[9] [47] é uma biblioteca *open source* em C++ que disponibiliza modelos pré-treinados para português para as seguintes tarefas de NLP: NER, atribuição de etiquetas morfossintáticas e divisão de tokens e frases.

Encontra-se disponível um manual de utilizador [10] completo e bem estruturado, no qual são descritos os procedimentos de instalação, importação para outras linguagens de programação, utilização de cada componente de NLP e o sistema de etiquetas, etc. Apesar de a maioria das tarefas estar disponível através da linha de comandos, algumas funcionalidades apenas são acessíveis usando a ferramenta como biblioteca.

[7] A documentação está acessível em https://polyglot.readthedocs.io/en/latest/Installation.html
[8] www.spacy.io
[9] NLP.lsi.upc.edu/freeling/index.php/
[10] O manual pode ser consultado em https://freeling-user-manual.readthedocs.io/en/latest/

4.1.6 Treetagger

Dada uma frase, Treetagger [48] revela informações morfossintáticas e o lema de cada palavra. Oferece modelos pré-treinados para diversas línguas, inclusive para português.

Os procedimentos relativos à sua instalação e uso são descritos de forma clara no site da ferramenta[11]. O modo de utilização de Treetagger não implica conhecimentos de programação, pois pode ser realizado através da linha de comandos. A ferramenta pode ser igualmente importada para Python por via de uma componente intermediária, como por exemplo, treetagger-python[12]. São também disponibilizados tutoriais[13] para as diferentes tarefas de NLP e nas linguagens de programação C++ e Python.

4.1.7 StanfordNLP

StanfordNLP[14] (ou Stanza) [49] é uma biblioteca para Python que permite realizar várias tarefas de NLP como a divisão de frases e tokens, a geração do lema das palavras, a análise morfossintática, NER e análise de dependências. Esta biblioteca oferece modelos pré-treinados para 53 línguas, inclusive para português. Na realidade, StanfordNLP é uma uma interface para Python da ferramenta Stanford CoreNLP, em Java, que o grupo *StanfordNLP*[15] disponibiliza. A interface para Python requer, conhecimentos em programação. No entanto, as mesmas funcionalidades que StanfordNLP apresenta podem ser executas por via da linha de comandos usando a ferramenta principal (Stanford CoreNLP).

Existem tutoriais[16] que descrevem os passos da instalação e utilização desta biblioteca, assim como exemplos para as diferentes tarefas de NLP, nomeadamente para a classificação de entidades nomeadas e análise morfossintática.

4.1.8 OpenNLP

OpenNLP[17] [50] é uma biblioteca open source para Java de processamento de texto baseada em aprendizagem automática e que fornece modelos pré-treinados, inclusive para texto em português para as tarefas: divisão de tokens, divisão de frases e atribuição de etiquetas morfossintáticas.

A documentação para a instalação da ferramenta não se encontra referenciada na página da ferramenta[18], o que dificultou a instalação. No entanto, no site da ferramenta existe um guia de referência bem estruturado, que descreve os modos de utilização da ferramenta, o procedimento para treino de

[11] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[12] https://github.com/miotto/treetagger-python
[13] https://freeling-tutorial.readthedocs.io/en/latest/
[14] stanfordNLP.github.io/stanfordNLP/
[15] Essas ferrramentas podem ser exploradas em https://NLP.stanford.edu/software/.
[16] Os tutorias podem ser consultados em https://stanfordNLP.github.io/stanfordNLP/installation_usage.html#getting-started .
[17] openNLP.apache.org/
[18] O procedimento da instalação encontra-se na página https://openNLP.apache.org/building.html

modelos e a execução de cada componente com base em modelos pré-treinados. Estas informações, são acompanhadas por exemplos. OpenNLP oferece um modo de utilização baseado na execução de programas na linha de comandos. Assim, para treinar, testar e aplicar esta ferramenta com modelos pré-treinados nas diferentes tarefas de NLP não são exigidos conhecimentos de programação. Por ser uma biblioteca direcionada para a linguagem de programação Java, a sua importação para Python requer uma componente que faça a ligação. Para tal, neste trabalho usou-se a interface nltk-openNLP[19].

4.1.9 Modelos pré-treinados do corpus SIGARRA NEWS

O INESC TEC[20] fornece modelos pré-treinados, doravante modelos-SIGARRA, direcionados para a classificação de entidades nomeadas em texto português europeu para as ferramentas OpenNLP, StanfordNLP, SpaCy e NLTK. Os corpora de treino desses modelos são compostos por 1000 artigos retirados da secção de notícias do sistema de informação da Universidade do Porto (SIGARRA).

4.2 Procedimento experimental

O primeiro passo foi criar um corpus de referência para as tarefas em análise. Esse corpus é descrito na Secção 4.2.1. O segundo passo foi atender a especificidades de cada ferramenta, o que levou a modificações do corpus de referência no que respeita à divisão em tokens feita pelas diferentes ferramentas (Secção 4.2.2). De seguida, foram ainda definidos vários *scripts* que transformam as etiquetas do corpus criado nas etiquetas usadas pelas diferentes ferramentas (Secção 4.2.3) e finalmente, apresentam-se as medidas de avaliação na Secção 4.3 .

4.2.1 Coleção dourada

A coleção dourada é composta por 101 frases retiradas de revistas, jornais e livros portugueses disponíveis *online*. Dessas frases, 59% pertencem a revistas como a Visão e a Exame Informática, 29% são de jornais (a maioria retiradas do jornal Observador e as restantes do Público) e 13% pertencem a um livro, como indica a Tabela A.1.

Antes da anotação das frases, foram corrigidos pequenos erros ortográficos pois este não era um ponto de avaliação. Assume-se assim que os textos que chegam às diferentes ferramentas estão limpos de erros ortográficos.

Seguidamente, foram anotadas as classes gramaticais e informações de flexão associadas a cada palavra. Por exemplo, para um substantivo, além do seu tipo (comum ou próprio), foram igualmente

[19] https://github.com/paudan/openNLP_python

[20] INESC TEC é o Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência e os modelos pré-treinados podem ser encontrados no site rdm.inesctec.pt/ro/dataset/cs-2017-006

anotados o seu género e número. Da mesma forma, para os verbos foram anotados o modo, o tempo, a pessoa e o número. Repetiu-se este processo para as outras classes gramaticais. a Tabela A.2 apresenta a frequência de cada classe e a Tabela A.3 lista a frequência das entidades nomeadas na coleção dourada.

4.2.2 Adaptação do corpus às diferentes ferramentas

Cada ferramenta apresenta as suas particularidades no que respeita ao modo como processa texto, o que implica que a coleção dourada anterior não possa ser usada sem antes ser pré-processada de acordo com as características de cada ferramenta. Por exemplo, a maior parte destas ferramentas não divide contrações (Polyglot, NLTK, SpaCy e OpenNLP). Por outro lado, a ferramenta FreeLing permite unir locuções e nomes compostos. Esta particularidade implica que, para esta ferramenta, sejam anotadas como uma unidade as locuções e nomes compostos da coleção dourada, ao invés de serem anotados separadamente os elementos gramaticais que compõem expressões.

A forma como as unidades textuais são segmentadas varia igualmente entre as diversas ferramentas. Por exemplo, algumas consideram "quarta-feira" como um um único token enquanto que outras separam cada elemento da palavra, ou seja, consideram "quarta", "−" e "feira" como três tokens distintos. Já o OpenNLP considera os verbos ligados a pronomes clíticos como um único token, como sucede em "reuniram-se" e em "escondê-lo". Esta variedade de características, tornou árdua a compatibilidade da coleção dourada principal com todas as ferramentas, pelo que foram criadas diferentes coleções douradas, específicas para cada ferramenta.

No que respeita à tarefa de NER, foram consideradas as entidades específicas de cada uma das ferramentas, de modo a que as saídas das diversas ferramentas fossem compatibilizadas com as entidades consideradas na coleção dourada.

4.2.3 Sobre as etiquetas morfossintáticas

Cada ferramenta tem o seu sistema de classes ou etiquetas morfossintáticas para as tarefas de NLP. O acesso a esta informação facilita a conversão automática das anotações da coleção dourada nas etiquetas das diferentes ferramentas. Esta tarefa não é, de todo, trivial devido às diferenças nas etiquetas das ferramentas e, em alguns casos, devido à indisponibilidade de um manual de utilizador que descrevesse o conjunto de símbolos possíveis. É o caso da ferramenta SpaCy, que não possui uma descrição das etiquetas morfossintáticas. No entanto, foi possível, na maioria dos casos, mapear sem problemas de maior, as etiquetas usadas pelo corpus de referência nas etiquetas das diferentes ferramentas. De notar ainda que algumas ferramentas contribuem com etiquetas com uma maior granularidade (por exemplo, o FreeLing, tal como se verá de seguida).

4.2.3.A NLTK, OpenNLP e NLPyport

O sistema de etiquetas do NLTK corresponde ao conjunto de etiquetas com a categoria gramatical utilizado na anotação dos corpora da Floresta Sintática[21]. Quanto ao OpenNLP, não se encontrou, na documentação da ferramenta, uma descrição do conjunto de etiquetas para o modelo em português. No entanto, verificou-se que o sistema de etiquetas do OpenNLP é semelhante ao da ferramenta NLTK, existindo uma ligeira diferença entre ambos no tratamento de sinais de pontuação: o OpenNLP agrupa os sinais de pontuação sob uma única etiqueta. Desta forma, para a conversão das anotações, procedeu-se de maneira semelhante à da ferramenta NLTK, à exceção dos sinais de pontuação que precisaram de um outro processamento. Por outro lado, as etiquetas consideradas nestas duas ferramentas não contêm informação de flexão. Além disso, o NLTK e o OpenNLP distinguem apenas verbos no infinitivo, gerúndio e particípio passado; todos as restantes formas verbais são classificadas de forma indiferenciada como "verbos finitos". Relativamente aos tipos de Pronomes só são identificados 3 tipos (determinativos, pessoais e independentes). A existência de uma descrição do conjunto de etiquetas facilitou a conversão das anotações da coleção dourada, apesar de não ser totalmente clara a categorização de alguns Determinantes e Pronomes, como por exemplo a classe pron-det (pronome determinativo). Por esta razão, as etiquetas relativas aos pronomes e determinantes serão tratadas à parte.

Dado que os recursos da ferramenta NLPyPort se baseiam na ferramenta NLTK e o modelo para esta tarefa foi treinado com os corpora Bosque da Floresta Sintática e Mac−Morpho [51], o conjunto de etiquetas morfossintáticas assemelha-se ao conjunto de etiquetas da ferramenta NLTK. Contudo, tal como no OpenNLP, os sinais de pontuação são agrupados numa mesma etiqueta, punc.

4.2.3.B Polyglot e StanfordNLP

As ferramentas Polyglot e StanfordNLP baseiam as suas etiquetas nas da CoNLL−U [52], compostas por vários elementos que descrevem morfossintaticamente uma palavra. Um dos elementos corresponde à classe gramatical universal, UPOS[22]; outro dos elementos vem do FEATS[23], que descreve informações morfológicas associadas à palavra. Com este conjunto universal de etiquetas, a correspondência entre as anotações da coleção de referência e as classes gramaticais realizou-se de uma forma mais simples. No entanto, o Polyglot só apresenta informações sobre a classe gramatical universal. Em particular, o conjunto de etiquetas morfossintáticas usado nos modelos do Polyglot corresponde às classes gramaticais principais da Tabela A.2, como adjetivo, determinante, advérbio, etc.

As etiquetas de StanfordNLP são compostas por vários campos indicando a palavra, o lema, a classe gramatical principal, aspetos morfológicos, entre outros. A etiqueta associada ao substantivo

[21] https://www.linguateca.pt/
[22] https://universaldependencies.org/u/pos/.
[23] https://universaldependencies.org/u/feat/index.html

comum “horas” é:

(15) *<Word index=20;text=horas;lemma=hora;upos=NOUN; xpos=_;feats=Gender=Fem|Number=Plur; governor=17;dependency_relation=obj>*

Para a tarefa de atribuição de etiquetas morfossintáticas, apenas as componentes UPOS (classe gramatical universal) e FEATS (informação morfológica) são interessantes para a avaliação. Além de o processamento realizado para isolar estas duas componentes, criou-se uma estrutura de etiqueta mais simples de se avaliar em relação à exposta em 15. Nesta nova estrutura as duas componentes FEATS e UPOS são separadas por |. A etiqueta do exemplo anterior, para a palavra “horas”, convertida para a nova estrutura corresponde simplesmente a etiqueta NOUN|Fem|Plur.

4.2.3.C SpaCy

No que respeita ao SpaCy, o sistema não disponibiliza um glossário completo com a descrição das etiquetas relativas à análise de texto em português, o que dificultou a compreensão dos resultados obtidos. Assim, o seu sistema de etiquetas não é claro, apesar de, aparentemente, se basear igualmente nos mesmos formalismos standard do Polyglot e StanfordNLP. No entanto, partilha as etiquetas relativas aos pronomes e determinantes do NLTK e OpenNLP. A conversão automática tornou-se mais complexa para esta ferramenta.

Tal como na ferramenta StanfordNLP, as etiquetas previstas foram processadas para se obter as informações relevantes.

4.2.3.D Freeling e Treetagger

No manual de utilizador da ferramenta Freeling encontram-se descritos os conjuntos de etiquetas morfossintáticas e de entidades nomeadas. Além das classes gramaticais principais, as etiquetas contêm informações sobre outros aspetos morfológicos como o género, número, modo e tempo verbal. Devido a esta especificação sobre a flexão nominal e verbal, o sistema de etiquetas desta ferramenta é extensivo, tornando a conversão automática complexa. Porém, a descrição clara de cada classe no manual de utilizador evitou dificuldades na correspondência entre as anotações da coleção dourada e as etiquetas.

O TreeTagger utiliza um sistema de etiquetas semelhante, no qual, além das classes gramaticais principais, são representadas outras informações morfológicas das palavras[24]. No entanto, as classes não possuem o mesmo nível de especificidade que as do FreeLing. Por exemplo, apenas a classe dos nomes incorpora informações sobre flexão em género e número. A descrição do conjunto das classes também é clara. Quanto às etiquetas dos verbos, só apresentam informação sobre o tipo

[24] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Portuguese-Tagset.html.

do verbo (auxiliar ou principal) e os modos verbais. Tal como na ferramenta FreeLing, as contrações são resolvidas automaticamente e identificadas com o sinal "+". Por exemplo, a contração "das" é identificada com "SPS + DA" por ser a contração de uma Preposição (SPS) com o Artigo Definido (DA) ("de + as").

4.2.4 Sobre as entidades nomeadas

Para a tarefa de NER, usaram-se os modelos pré-treinados para as ferramentas SpaCy, NLTK, OpenNLP e StanfordNLP, descritos em [53]. Devido a dificuldades no carregamento do modelo para a ferramenta SpaCy, talvez por incompatibilidade da versão mais recente da ferramenta com o modelo, este não foi avaliado. De acordo com [53], o seu conjunto de etiquetas corresponde a oito classes relacionadas com o domínio dos corpora: Hora, Evento, Organização, Curso, Pessoa, Localização, Data e UnidadeOrganica.

No que respeita às restantes ferramentas, o FreeLing apresenta as seguintes classes de entidades e as respetivas etiquetas: Pessoa (NP00SP0), Localização (NP00G00), Organização (NP00O00) e Outros (NP00V00). Esta última etiqueta corresponde a entidades nomeadas que não se integram em nenhuma das categorias anteriores. No entanto, por ser uma classe ambígua, tokens que não são reconhecidas como entidades na coleção dourada como "Verão" são classificados como Outros. Por outro lado, entidades nomeadas identificadas na coleção dourada como Outros são classificados como Organização. Além disso, entidades nomeadas de outras classes como "ArsTechnica" são consideradas como Outros. Para facilitar a avaliação automática, foi adicionada mais uma etiqueta ("O") que identifica os tokens que não são entidades. Finalmente, quanto ao Polyglot, este deteta apenas 3 classes de entidades: Pessoa (I-PER), Localização (I-LOC) e Organização (I-ORG). Tal como para a ferramenta anterior foi adicionada a etiqueta "O", com o mesmo significado.

4.2.5 Medidas de avaliação

Na avaliação das ferramentas usou-se a Micro− e a Macro−Média da medida F1, tal como descritas por [54] e implementadas no SCIKIT-LEARN[25].

4.3 Resultados da análise morfossintática

Nesta secção começa-se por discutir alguns aspectos relativos aos modelos criados para a ferramenta NLTK. De seguida, apresenta-se e discute-se os resultados obtidos pelos diferentes modelos e ferramentas.

[25] https://scikit-learn.org

4.3.1 Sobre o Modelos do NLTK

Todas as ferramentas com a excepção do NLTK oferecem modelos pré-treinados. Assim, para esta ferramenta, foram treinados vários modelos. Um desses modelos é o modelo de Bigramas, tal como apresentado nos manuais do NLTK. Foram ainda implementados modelos baseados na Máxima Entropia e ainda um modelo denominado Perceptrão, pois o OpenNLP fornece modelos baseados nestes dois algoritmos e considerou-se interessante a sua comparação. De notar que poderiam ter sido implementados modelos mais adequados à predição de sequências, como os HMM [55] ou os CRF [56] ou ainda modelos baseados em redes neuronais. No entanto, a ideia foi testar o que a ferramenta oferece directamente. Todos estes modelos foram treinados nos corpora "Floresta Sintática", disponíveis com o NLTK[26]. De notar que a versão disponível no NLTK da Floresta Sintática não contém alguns sinais de pontuação como dois-pontos $<:>$, reticências $< \cdots >$, parênteses curvos e o símbolo de percentagem $< \% >$, pelo que estes items não são previstos pelos diferentes modelos. A divisão em tokens das frases foi realizada pela função *word_tokenize*[27]. A criação de qualquer um dos modelos não requer um grande esforço na ótica de um informático, pois consiste em importar uma biblioteca própria para o efeito. No entanto, alguns modelos têm as suas especificidades. Assim, para o modelo de bigramas, cada bigrama é um *token*, tal como identificado pelo *tokenizador*; no caso de aparecer um *token* nunca visto no treino, decidiu-se atribuir, por omissão, a etiqueta "n" (classe gramatical Nome Comum), por ser a classe mais comum. Nos casos em que os modelos requeriam a recolha de características (*features*) dos dados, foram usadas características muito simples, como a própria palavra, a palavra anterior e a seguinte, se a palavra corrente começava com maiúscula, etc. A seleção das *features* dita o desempenho da ferramenta. No entanto, está fora do âmbito deste trabalho desenvolver um estudo exaustivo sobre as *features* a utilizar.

4.3.2 Resultados Globais

A Tabela B.1 apresenta os resultados globais dos diferentes modelos, tendo em conta a totalidade das etiquetas (de cada ferramenta) e considerando a Micro- e a Macro-média relativas à medida F1. Devemos realçar que estes valores não permitem uma comparação totalmente justa dos diferentes modelos, pois, como dito anteriormente, os valores de Micro- e Macro-Média da F1 são calculados com base no conjunto de etiquetas de cada ferramenta, que varia entre estas no que diz respeito à informação de flexão. Assim, estes valores devem dar apenas uma ideia geral. De modo a comparar de forma justa estes modelos, a secção seguinte detalha os valores para estas medidas sem ter em conta as informações de flexão.

[26] Os modelos disponibilizados pelo OpenNLP, SpaCy e StanfordNLP também utilizam este corpus.

[27] Esta função tem a particularidade de mudar as aspas iniciais $<$"$>$ de uma frase para dois acentos graves $<``>$ e as finais $<$"$>$ para duas plicas $<''>$. Por isso, é necessária uma reconversão para aspas, antes do treino e teste dos dados.

4.3.3 Resultados por Classe Gramatical

A Tabela B.2 permite comparar os diferentes modelos tendo em conta as categorias que têm em comum e sem ter em conta as informações de flexão. De notar que são apenas mostradas as classes gramaticais mais relevantes ('Bg' representa o modelo baseado em Bigramas, 'P' o Perceptrão, 'ME' a Máxima Entropia, 'NLPy' o NLPyPort, 'PG' o Polyglot, 'TT' o TreeTagger, 'FL' o FreeLing, 'StfNLP' o StanfordNLP e 'ONLP' o OpenNLP). As Tabelas B.3 e B.4 apresentam os resultados para os dois grupos de modelos que partilham algumas etiquetas específicas.

4.3.4 Discussão

Os melhores resultados globais são obtidos pelo OpenNLP, Máxima Entropia, 91% de Macro-Média e aos 94% de Micro-Média, porém, é importante realçar que o nível de detalhe do conjunto de etiquetas da ferramenta não é tão fino quanto nos modelos anteriores, pois, tal como referido anteriormente, o conjunto de etiquetas do OpenNLP não contém informações de flexão de número, género e tempos verbais (ao contrário de outras ferramentas como o FreeLing, o StanfordNLP e o TreeTagger, que apresentam etiquetas mais finas).

Quanto à avaliação por classe, a ferramenta StanfordNLP é aquela que apresenta maiores valores na previsão de verbos no conjuntivo e no condicional. O FreeLing apresenta melhor desempenho na classificação de adjetivos, nomes próprios e conjunções subordinativas. Por fim, o TreeTagger destaca-se na previsão de advérbios (de notar que nesta ferramenta o condicional não é considerado modo mas sim um tempo verbal do modo indicativo).

Existem ainda algumas particularidades das ferramentas, que merecem ser discutidas. Por exemplo, como dito anteriormente, o FreeLing reconhece nomes compostos como um único *token* (por exemplo, "Eduardo Cabrita" é tratado como um *token* único "Eduardo_Cabrita"). Quando identifica sequências de palavras que formam expressões compostas, devolve-as unidas (por exemplo, a locução conjuncional "mesmo que" é reconhecida como o único token "mesmo_que"). Estas particularidades auxiliam a classificação de nomes próprios compostos, evitando assim que preposições e nomes próprios sejam incorretamente classificados, o que é constante nas outras ferramentas. Desta forma, o desempenho da previsão de nomes próprios alcançou o valor maior de F1-measure (96%). Infelizmente esta ferramenta não classifica corretamente verbos auxiliares (VA). Neste aspecto, a melhor ferramenta é o SpaCy (84%), apesar de o StanfordNLP também conseguir identificar verbos auxiliares (73%). O FreeLing tem também outras características que o tornam interessante: palavras relacionadas com datas são unidas e consideradas como um só token, à semelhança dos nomes próprios, como "Novembro de 2018", que é tratado como "Novembro_de_2018". O critério de atribuição desta classe é, contudo, pouco claro, por não haver uma descrição desta etiqueta. Esta ferramenta também une numerais, por exemplo,

considera “10 mil milhões” como um único *token*. Estas uniões não são totalmente precisas, levando a erros de classificação de alguns *tokens*, como acontece na expressão “15 e 35”, que é considerado como um só *token* “15_e_35”. Esta situação leva, posteriormente, a uma incorreta classificação, neste caso atribuindo a classe Interjeição.

4.4 Resultados do reconhecimento de entidades nomeadas

Nesta secção serão apresentados os resultados e conclusões sobre o desempenho das ferramentas na tarefa de NER.

4.4.1 Sobre os modelos-SIGARRA

No que diz respeito ao NLTK, tal como previamente indicado, estudaram-se três modelos-SIGARRA, pré-treinados no corpus SIGARRA NEWS: modelo baseado em Árvores de Decisão, no Naïve Bayes e na Máxima Entropia. Os modelos avaliados associados ao OpenNLP e StanfordNLP são também os referidos modelos-SIGARRA, treinados no mesmo corpus, mas com estas ferramentas.

4.4.2 Resultados Globais

A Tabela B.5 mostra os resultados globais dos diferentes modelos, tendo em conta a Micro- e a Macro-Média relativas à medida F1. O FreeLing é o melhor sistema quanto à Micro-Média e o StanfordNLP quanto à Macro-Média. De notar que, sendo os modelos do NLTK, StanfordNLP e OpenNLP, modelos-SIGARRA, há uma diferença substancial de valores quanto à Macro-Média (o modelo Naïve Bayes com 0.18 e o StanfordNLP 0.78), assunto que se discutirá mais à frente.

4.4.3 Resultados por Tipo de Entidade

A Tabela B.6 apresenta a comparação entre os valores de F1-measure de cada classe e para cada ferramenta na tarefa de NER. O StanfordNLP é a ferramenta os melhores resultados em praticamente todas as classes, à exceção da classe Localização, na qual a ferramenta Freeling se destaca.

4.4.4 Discussão

Como dito anteriormente, há uma grande diferença de valores quanto à Micro- e Macro-Média em alguns modelos. Na verdade, os resultados apresentados mostram quão enganadora pode ser a Micro-Média (F1). Esta medida tem em conta a soma de todos os Verdadeiros/Falsos Positivos/Negativos de todas as classes, sendo calculada posteriormente a Precisão e a Cobertura, e, finalmente, a F1. Ora,

todos os casos que não são considerados como entidades mencionadas e não são de facto entidades mencionadas (isto é, o grosso das palavras, pois a maioria das palavras de um texto não são entidades mencionadas) contam como Verdadeiros Positivos, indicando que esta métrica não é nada informativa neste cenário em que as classes não são balanceadas.

Por outro lado, e como referido, os algoritmos usados têm um papel extremamente relevante na tarefa de NER. Os resultados dos diferentes algoritmos na base dos modelos-SIGARRA treinados com o NLTK ilustram bem essa situação.

É também interessante verificar como a estratégia de *tokenização* usada pelas ferramentas pode fazer a diferença. No caso do FreeLing, a divisão em tokens desta ferramenta foi um alicerce na classificação correta de entidades nomeadas, pelo simples facto de preservar nomes compostos, tornando possível a identificação de entidades compostas como "Parques de Sintra-Monte da Lua", "Estados Unidos da América" e "Diário de Notícias". Outra característica que se realça corresponde à sua capacidade para identificar entidades nomeadas estrangeiras como "Sujoy Ghosh", "Einstein" e "Xuekun Fang". Em contrapartida, verificou-se que alguns *tokens* presentes no início das frases foram incorretamente consideradas como entidades, talvez por começarem com letra maiúscula. Segue-se um exemplo desse caso: "Contactada pela Lusa. . . ".

A análise da classificação do Polyglot revelou que esta considera nacionalidades tais como "mexicanos", "dinamarquesa" e "americanos" sempre como Localização. No entanto, esta ferramenta consegue identificar entidades nomeadas estrangeiras como "Einstein", "Kaiserslautern" e "Sujoy Ghosh" e algumas entidades nomeadas compostas como "Estados Unidos", "Universidade de Aveiro" e "Physical Review". Já o OpenNLP tem dificuldades a identificar entidades nomeadas estrangeiras, como por exemplo, "Netflix", "Maximilian Gunther", "Einstein"; por outro lado, não é totalmente capaz de classificar nomes compostos. Por exemplo, "Associação de Proteção e Socorro" e "Universidade de Londres" são corretamente identificadas. Contudo, outras entidades como "Universidade da Califórnia", "Diário de Notícias", "Estados Unidos" e "Eduardo Cabrita" já não o são. Outra conclusão retirada é que não identifica siglas como "EUA", "MIT", "PSML", "EHT", etc.

O StanfordNLP, por seu turno, consegue identificar entidades nomeadas noutras línguas, incluindo siglas. No entanto, verificou-se que, quando uma sigla está entre parênteses, a ferramenta identifica os parênteses como fazendo parte da entidade nomeada. Por exemplo, dada a sigla MIT neste formato "(MIT)", além de MIT ser considerada como Organização, o parêntese "( )" também o é. De entre as ferramentas analisadas para esta tarefa, o StanfordNLP apresenta um melhor desempenho em praticamente todas as classes, à excepção das classes Localização e Organização, onde o vencedor é o FreeLing.

5

PE2LGP 4.0

Conteúdo

PE2LGP 4.0 é o sistema de tradução implementado neste trabalho. Integra-se no projeto Corpus linguístico & Avatar da língua gestual portuguesa do Instituto de Ciências da Saúde da Universidade Católica Portuguesa, em parceria com o INESC-ID e financiado pela Fundação para a Ciência e a Tecnologia (FCT) (PTDC/LLTLIN/29887/2017).

O sistema de tradução divide-se em dois módulos principais, como indicado na Figura 5.1. O primeiro módulo (Secção 5.1), *Construção de regras de tradução*, consiste na extração de informações linguísticas do corpus de referência (Secção 2.3) e, a partir dessas informações, na criação de regras automáticas e de um dicionário bilingue de português e LGP. O segundo módulo (Secção 5.2), *Tradução automática*, consiste na tradução de texto em PE para LGP, em que a frase em LGP é representada por uma sequência de glosas com marcadores que identificam as expressões faciais e palavras soletradas. Na base da tradução encontram-se as regras automáticas e o dicionário bilingue criados no primeiro módulo e ainda regras manuais que capturam fenómenos linguísticos relacionados com a morfologia das palavras, como a marcação do feminino e outros que as regras automáticas não cobrem, como as interrogativas parciais[1] e expressões faciais. Estes dois módulos possuem uma componente comum, a fase de análise. Nas próximas secções as componentes de cada módulo serão detalhadas.

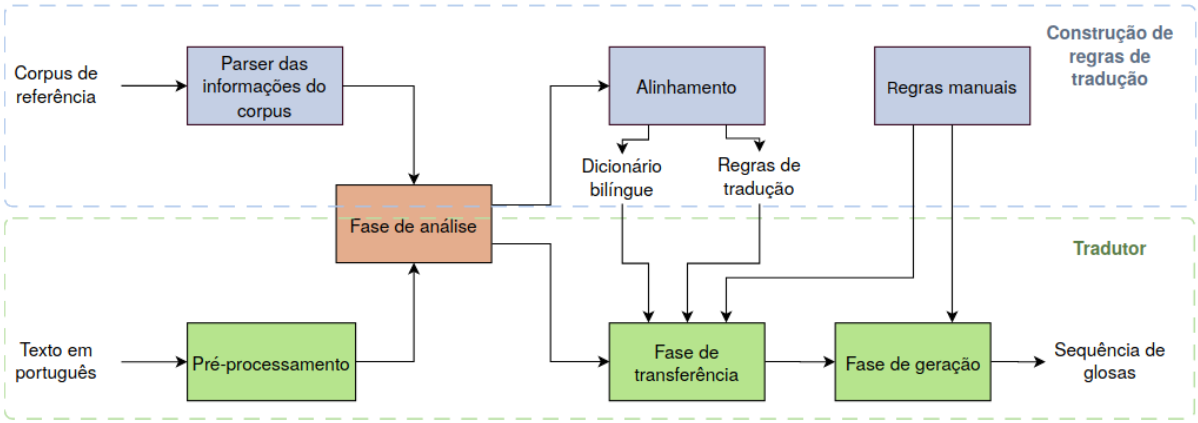


Figura 5.1: Arquitetura do sistema de tradução PE2LGP 4.0.

5.1 Módulo de construção das regras de tradução

A falta de estudos linguísticos sobre a LGP é um obstáculo para a criação de um tradutor baseado em regras manuais que possa reproduzir de forma fiel os seus fenómenos linguísticos. Nesta tese esse obstáculo é contornado por este módulo, onde são extraídas informações gramaticais do corpus de referência da Universidade Católica Portuguesa (Secção 2.3) e construídas, a partir dessas informações,

[1] De acordo com a página ciberdúvidas interrogativas parciais possuem um elemento interrogativo (QU-), no qual, fazem parte os pronomes e advérbios interrogativos, como Quem e Onde.

regras que descrevem como uma frase em português pode ser produzida em LGP, doravante conhecidas como **regras de automáticas** (Secção 5.1.5), bem como um dicionário bilingue de português-LGP (Secção 5.1.7). A essas regras automáticas acrescenta-se um conjunto de regras manuais[2] (Secção 5.1.8) construídas a partir das características gramaticais da LGP listadas na Secção 2.

O corpus de referência detalhado na Secção 2.3 é composto pelas transcrições dos enunciados dos vídeos e pelas suas traduções em português e informações gramaticais das frases em LGP, nomeadamente as classes gramaticais e os argumentos externos e internos (sujeitos e objetos). Estas anotações são realizadas no ELAN[3]. Deste modo, o primeiro passo consiste na exportação e processamento dessas informações (Secção 5.1.2). Dado que as regras automáticas refletem aspetos gramaticais das duas línguas e, do ELAN são somente exportados os aspetos gramaticais das frases em LGP, na fase de análise (Secção 5.1.3) as frases em português são analisadas para se conhecerem as suas informações gramaticais. Tendo as informações das frases de ambas as línguas, procede-se ao alinhamento entre as palavras e os gestos das frases (Secção 5.1.4). Dos pares palavra-gesto alinhados e das suas informações gramaticais, constroem-se as regras automáticas (Secçao 5.1.5) e o dicionário bilingue (Secção 5.1.7). As regras automáticas dividem-se em dois tipos: as que descrevem a ordem dos constituintes morfossintáticos ou estrutura sintática (doravante **regras morfossintáticas**) e as que descrevem a ordem frásica (**regras frásicas**). As primeiras regras são agrupadas por **elemento frásico**, ou seja, constroem-se regras para os modificadores de frase, regras para o sujeito e regras para o predicado. A arquitetura deste módulo encontra-se na Figura 5.1, delimitada pelo retângulo azul (o retângulo em cima).

5.1.1 Sobre os dados usados

Os dados utilizados neste módulo provêm de um vídeo de 5 minutos de um gestuante nativo. O tipo de discurso do vídeo é informal e espontâneo. O tema do enunciado é a *história do Barroco*. Os dados extraídos dizem respeito à transcrição do enunciado gestuado e à sua tradução para português europeu e às **informações gramaticais** das frases em LGP, nomeadamente as classes gramaticais, a análise sintática e o tipo da frase (Interrogativa (INT), Negativa (NEG) e Exclamativa (EXCL)). No corpus não foram anotados aspetos gramaticais da frase em português. Seguem-se algumas características desses dados:

- Das 66 frases que constituem estes dados, 3 são declarativas negativas, 5 interrogativas e as restantes declarativas afirmativas.
- As frases em LGP são transcritas em glosas.

[2]Neste documento, as regras automáticas são as regras extraídas do corpus e as regras manuais, aquelas que foram construídas manualmente com base nas diferenças gramaticais conhecidas entre as duas línguas.

[3]É uma ferramenta que permite a criação de várias camadas de anotações de vídeos e áudio. Pode ser acedida em tla.mpi.nl/tools/tla-tools/elan

- As classes gramaticais das frases em LGP pertencem ao conjunto das classes principais (determinantes, pronomes, adjetivos, substantivos, etc.), sem identificação da subclasse.

- A análise sintática identifica os **argumentos externos** (sujeitos), **internos** (objetos) e verbos (transitivos, intransitivos, copulativos, etc.) de uma frase em LGP.

O corpus ainda está em desenvolvimento e novas regras podem ser geradas à medida que o corpus vai crescendo.

5.1.2 Exportação e *parse* das informações do *ELAN*

Os dados anteriores são exportados do *ELAN* num ficheiro *HTML*, que é processado pela componente *parser*, obtendo-se as frases em português, as frases em LGP e as suas informações gramaticais.

5.1.3 Fase de análise

As informações gramaticais extraídas do corpus de referência dizem respeito apenas às frases em LGP, mas para construir automaticamente as regras são igualmente necessárias as informações gramaticais das frases em português. Obter estas informações é o principal propósito desta fase, e são conseguidas analisando sintática e morfossintaticamente as frases em português com recurso a ferramentas de NLP (Secção 5.1.3.A). Na análise sintática (Secção 5.1.3.C), identificam-se os **elementos frásicos** (sujeito, predicado e modificador de frase) e na análise morfossintática (Secção 5.1.3.B), conhecem-se as classes gramaticais das palavras (ou constituintes morfossintáticos).

5.1.3.A Ferramentas de NLP

O desempenho e as características das ferramentas discutidos no Capítulo 4, permitiram escolher a ferramenta que melhor se ajusta às necessidades do sistema de tradução. Para a análise morfossintática escolheu-se a ferramenta Freeling e para a análise sintática, o SpaCy.

5.1.3.B Análise morfossintática

As classes e subclasses gramaticais (determinantes possessivos, determinantes demonstrativos, etc.) bem como aspetos de flexão (em género, número, tempo verbal e modo verbal, etc.) e os lemas das palavras das frases em português (e dos gestos das frases em LGP) são conhecidos nesta etapa através das análises sintática e morfossintática. Este último passo é realizado por ser a base do alinhamento de palavras e gestos descrito na Secção 5.1.4.

5.1.3.C Análise sintática

As relações de dependência entre os constituintes de uma frase foram identificadas pela ferramenta SpaCy. Estas relações formam uma estrutura hierárquica entre os constituintes, podendo ser representadas num grafo (grafo de dependências), que é uma das representações utilizadas pela ferramenta.

Por exemplo, as relações entre os constituintes da frase *A Maria comeu um gelado.* formam o grafo de dependências na Figura 5.2.

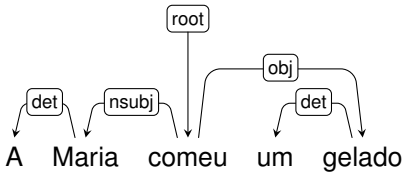


Figura 5.2: Grafo de dependências da frase *A Maria comeu um gelado.*

Neste exemplo, o verbo é o elemento principal e por isso é a raiz do grafo de dependências. Contudo, nem sempre se verifica isso. Quando o verbo é copulativo, como na frase *Infelizmente, o Miguel está constipado.*, o elemento principal é o predicativo do sujeito, *constipado*. O verbo neste caso, exerce uma função de ligação entre o sujeito e uma característica do sujeito. O grafo de dependências deste exemplo está representado na Figura 5.3.

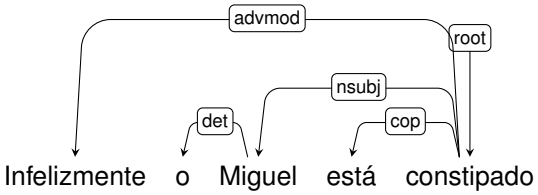


Figura 5.3: Grafo de dependências da frase *Infelizmente, o Miguel está constipado.*

Os elementos que são filhos da raiz do grafo de dependências representam os **elementos frásicos** (sujeito, objeto e modificador de frase). A partir destas informações sabe-se a ordem frásica. Para a frase representada na Figura 5.2, *a Maria* é o sujeito (nsubj) e *um gelado* é o objeto (obj) e consequentemente, a ordem frásica é SVO. A frase do grafo na Figura 5.3 só tem sujeito (*o Miguel* (nsubj)) e modificador de frase (*Infelizmente* (advmod)) e, por isso, a sua ordem frásica é sujeito-verbo (SV).

Além das etiquetas *nsubj*, *obj*, *advmod*, outras poderão surgir. O mapeamento entre as possíveis etiquetas e o respetivo elemento frásico (sujeito, predicado ou modificador) foi realizado com base nas definições e exemplos das mesmas, que podem ser consultados em `universaldependencies.org/u/dep/index.html`.

5.1.3.D Pós-processamento

Dado que a LGP não possui determinantes artigos definidos e indefinidos, estes foram removidos da frase, assim como a pontuação. As preposições foram igualmente eliminadas por não serem representadas em LGP isoladamente (ver Secção 2.2.8). Este fenómeno não pertence ao âmbito deste trabalho.

As etiquetas resultantes da análise morfossintática são mapeadas para as etiquetas do corpus para uniformizar as etiquetas das regras. Por exemplo, a etiqueta *NCMS000* da ferramenta FreeLing refere-se a um nome comum no singular e no género masculino, e é convertida para *N*, de acordo com as convenções do corpus expostas na Tabela 2.1. Por sua vez, as etiquetas da análise sintática da ferramenta SpaCy e as do corpus são convertidas para uma notação mais simples; por exemplo, as etiquetas referentes a sujeitos são convertidas para *S* e as que identificam objetos são renomeadas para *O*.

Assim, da fase de análise conseguem-se as classes morfossintáticas, a ordem frásica da frase em português e os seus elementos frásicos, completando as informações necessárias para a construção automática das regras de tradução.

5.1.4 Alinhamento do corpus

Antes de passar à construção da gramática, há que alinhar o léxico (palavras e gestos) das frases do corpus. As correspondências entre uma palavra e um gesto não são simplesmente um-para-um. A Figura 5.4 esquematiza outros tipos de relações possíveis no alinhamento. A última relação é difícil de identificar, pelo que o seu tratamento foi excluído do âmbito do trabalho.

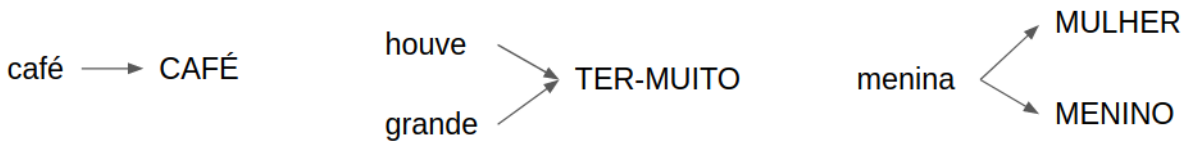


Figura 5.4: Os diferentes tipos de alinhamentos entre palavras e glosas. A primeira representa uma correspondência um-para-um, a segunda muitos-para-um e a última, um-para-muitos.

O treino do alinhamento de palavras e frases de um corpus é a base dos sistemas de tradução estatísticos (Secção 3.2.2) [40, 57, 58]. Inclusive, existe software capaz de alinhar automaticamente palavras em corpora, como o *Giza ++*[4]. No caso do par de línguas português-LGP, não existe um corpus suficientemente grande que permita criar um modelo probabilístico de alinhamento de palavras e gestos. Desta forma, desenvolveu-se um algoritmo de alinhamento baseado em medidas de semelhança, *string matching* e semelhança semântica.

[4] Pode ser obtido em https://github.com/moses-smt/giza-pp

O alinhamento é composto por 3 principais etapas. Primeiro, a palavra e o gesto são comparados letra-a-letra (Secção 5.1.4.B), se forem iguais, são alinhados. Caso contrário, as palavras e gestos são comparados com base no seu significado, recorrendo primeiro à *Wordnet* (Secção 5.1.4.C) e depois aos *word embeddings* (Secção 5.1.4.D). Esta última etapa vem reforçar o alinhamento semântico, alguns pares palavra-glosa que não são alinhados pela wordnet, poderão sê-lo com word embeddings. Nas próximas secções estas etapas são detalhadas. O pseudo-código do alinhamento está descrito no Anexo C.1.

De notar que o alinhamento é realizado por elemento frásico, ou seja, as palavras do predicado da frase em português são alinhadas com os gestos do predicado da frase em LGP. Por exemplo, para as frases em 16 e 17, o alinhamento é realizado separadamente entre os termos dos sujeitos: *A Maria e o João* e *MARIA JOÃO* e os termos dos predicados: *vão à piscina* e *PISCINA IR*.

(16) A Maria e o João vão à piscina.

(17) MARIA JOÃO PISCINA IR

Antes do alinhamento, os gestos e as palavras foram convertidos nos seus lemas e as glosas para minúsculas. Estes processamentos permitem alargar o número de correspondências exatas apanhadas pela primeira etapa (Secção 5.1.4.B). Por exemplo, os verbos *vão* e *IR*, não seriam considerados correspondências na primeira etapa, mas com este processamento ambos são convertidos para *ir*, tornando-se numa correspondência exata.

De seguida, as várias etapas do alinhamento são detalhadas e exemplificadas com base nos seguintes predicados:

(18) haver grande desenvolvimento artístico económico religioso político e social

(19) TER-MUITO ARTE IGREJA POLÍTICA SOCIAL DINHEIRO DESENVOLVIMENTO

5.1.4.A Decomposição de gestos compostos

Existem gestos representados por mais do que uma palavra em português (chamados de **gestos compostos**), como o gesto *TER-MUITO* na frase em 19. Estes gestos estão associados a correspondências de muitos-para-um, nas quais um gesto corresponde a mais do que uma palavra. Para poderem ser alinhados com os respetivos lemas, são previamente decompostos. O gesto *TER-MUITO* é decomposto nas glosas *TER* e *MUITO*, que serão alinhadas respetivamente com os lemas *haver* e *grande*.

5.1.4.B Lema e glosa iguais

Esta corresponde à primeira etapa do alinhamento. Dado que as glosas são um sistema de escrita das línguas gestuais baseado nas línguas orais, a maioria dos gestos são representados pela palavra

equivalente em português. Por exemplo, o gesto para *desenvolvimento* pode ser representado pela glosa *DESENVOLVIMENTO*. Assim, a primeira etapa alinha lemas e glosas iguais, alinhando a maioria das palavras e gestos. Os lemas e as glosas que não são alinhados aqui, são comparados a nível semântico nas próximas etapas.

Com este passo alinham-se quatro pares lema-glosa das frases em 18 e 19: século-SÉCULO, 17-17, social-SOCIAL e desenvolvimento-DESENVOLVIMENTO, correspondendo a metade dos termos das frases.

5.1.4.C *WordNet*

Wordnets são bases de dados lexicais de relações semânticas entre palavras, como relações de sinonímia, hiperonímia e holonímia. Os sinónimos que expressam um mesmo conceito são agrupados em *synsets*, que se interligam através de relações de hiperonímia e de hiponímia. A Figura 5.5 apresenta um exemplo simples de uma wordnet relacionada com o conceito *animal*. Os nós do grafo do exemplo correspondem a synsets, ou seja, animal de estimação, animal doméstico e animal de companhia são sinónimos. Por sua vez, este synset é hiperónimo dos synsets *cão e cachorro* e *gato*.

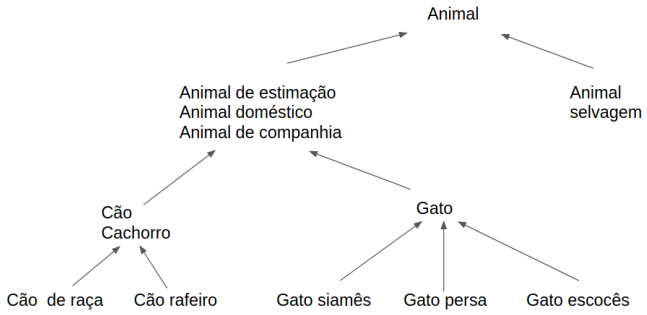


Figura 5.5: Exemplo de uma hierarquia numa wordnet.

Hugo Oliveira, Valeria de Paiva et al. [59] realizaram um levantamento das existentes ontologias lexicais disponíveis para português e das suas características, como as abordagens de criação e licenças de utilização de cada uma e a dimensão e cobertura das *wordnets*. De acordo com os autores, para o léxico português estão gratuitamente disponíveis as seguintes wordnets: *Onto.PT*[5], a *PULO*[6] e a *OpenWordNet-PT*[7].

A OpenWordNet-PT, por estar integrada na biblioteca NLTK, foi a usada neste trabalho para calcular a semelhança semântica entre um lema e um gesto. A biblioteca oferece várias medidas de semelhança entre dois conceitos. Uma delas é a semelhança de Wu-Palmer[8], que calcula uma pontuação que mede

[5] Disponível em `ontopt.dei.uc.pt/index.php?sec=projecto`.
[6] Disponível em `wordnet.pt/`.
[7] Acessível em `github.com/own-pt/openWordnet-PT`.
[8] Descrita em `www.nltk.org/howto/wordnet.html`.

a semelhança entre os significados de duas palavras com base na sua posição e significado indicados na Wordnet.

Considerou-se que um lema e um gesto são semanticamente semelhantes se possuírem um par de sinónimos com valor de semelhança de Wu-Palmer maior ou igual a 0.9. Contudo esta premissa revelou-se insuficiente, dado que estas medidas de semelhança são apenas válidas entre conceitos com a mesma classe gramatical [60]. Por exemplo, não é possível comparar com esta medida, os conceitos *arte* e *artístico* por serem de classes diferentes, o primeiro é nome comum e o segundo, adjetivo. Esta limitação poderia ser contornada se houvesse uma ferramenta que automaticamente derivasse uma palavra nas suas diferentes formas (nominal, adjetival, verbal, etc.). Não havendo, adicionou-se outra premissa: um lema e um gesto são também semanticamente semelhantes se possuírem sinónimos com radicais semelhantes, como as palavras ***arte*** e ***art****ístico*. Assim, para os pares de sinónimos com diferentes classes gramaticais e para aqueles com valor de semelhança anterior menor do que 0.9, calculou-se a Distância de Jaro-Winkler[9]. Com esta medida, conceitos com prefixos comuns são mais semelhantes.

Se para um lema e uma glosa existir um par de sinónimos com valor de Distância de Jaro-Winkler maior do que 0.8, então, esse lema e essa glosa são alinhados. Caso contrário, passa-se para a etapa seguinte, onde são comparados usando *word embeddings*.

5.1.4.D *Word embeddings*

Word embeddings são modelos estatísticos que permitem representar palavras ou frases em vetores de números de acordo com o contexto em que as palavras aparecem [61].

Nathan S. Hartmann, Erick Fonseca et al. [61] avaliaram 31 modelos[10] de word embeddings para português do Brasil e europeu, treinados com vários algoritmos como, FastText, GloVe, Word2Vec, etc. A avaliação revelou que para a analogia semântica e para o português europeu, o modelo com melhor desempenho é o treinado com o algoritmo GloVe com 600 dimensões. Este modelo converte o lema e a glosa em vetores. A semelhança entre as duas palavras relaciona-se com o ângulo formado pelos seus vetores: quanto menor for o ângulo entre os vetores, maior é a semelhança entre as palavras. Assim, a semelhança entre os vetores é calculada através de a Similaridade do Cosseno[11]. Se o lema e a glosa tiverem um valor de semelhança maior do que 0.3, então são alinhados.

Os limites de semelhança usados nos passos anteriores foram decididos com base nos resultados das diferentes medidas de semelhança aplicadas a 36863 pares de lemas e gestos, retirados de 4 vídeos do corpus de referência.

[9] A biblioteca pyjarowinkler para Python foi usuada para o cálculo da Distância de Jaro-Winkler.
[10] Encontram-se disponíveis em `nilc.icmc.usp.br/embeddings`.
[11] Detalhes sobre esta medida podem ser encontrados em `www.sciencedirect.com/topics/computer-science/cosine-similarity`.

No final do alinhamento tem-se, para o exemplo, os seguintes pares lema-gesto alinhados: século-SÉCULO, 17-17, haver grande-TER-MUITO, artístico-ARTE, religioso-IGREJA, político-POLÍTICA, SOCIAL-SOCIAL e desenvolvimento-DESENVOLVIMENTO. Este algoritmo implementado não forma um alinhamento perfeito, alguns pares não são alinhados, como o par económico-DINHEIRO.

5.1.5 Regras de tradução automáticas

Neste módulo são construídos dois tipos de regras automáticas: as que descrevem a estrutura sintática de cada elemento frásico (*regras morfossintáticas*) e as que descrevem a ordem frásica (*regras frásicas*). Estas regras são construídas a partir do alinhamento das palavras com os gestos das frases e das informações inferidas pela fase de análise relativas à frase em português (na Secção 5.1.3).

As ordens frásicas e dos constituintes morfossintáticos podem ser alteradas conforme o tipo de frase. Este fenómeno acontece noutras línguas, como no inglês, em que o sujeito nas frases interrogativas aparece depois do verbo auxiliar, ao contrário das frases declarativas, nas quais, normalmente, o sujeito aparece antes dos verbos. Por esta razão, as regras automáticas são agrupadas de acordo com o tipo da frase (declarativa afirmativa, negativa, interrogativa e exclamativa) que originou a regra.

As regras automáticas descrevem as transformações gramaticais necessárias para que uma frase em português possa ser convertida na frase em LGP. Assim, as regras são compostas por dois lados, pelo **lado português** e o **lado da LGP**. Os exemplos de regras dados daqui em diante seguem a estrutura em 20, em que o lado português e o lado da LGP são divididos por uma seta.

(20) lado português $\rightarrow$ lado da LGP

As regras frásicas construíram-se a partir das ordens frásicas de cada frase em português, dadas pela análise sintática (Secção 5.2.4) e da ordem frásica da respetiva frase em LGP extraída do corpus. Antes da definição das regras, as convenções usadas para identificar o sujeito, o objeto e verbo no corpus de referência são convertidas para a nomenclatura usada na análise sintática. Por exemplo, *ARG_EXT* que identifica o sujeito da frase é convertido para *S*, o *ARG_INT* (objeto) passa a ser representado por *O* e todos os verbos (verbos transitivos, *V_TRAN*, e intransitivos, *V_INTR*, etc.) são identificados por *V*. Um exemplo de uma regra frásica construída a partir das informações de uma frase interrogativa está em 21.

(21) SVO $\rightarrow$ SOV

A construção das regras morfossintáticas baseia-se nas classes gramaticais dos elementos que compõem os pares lema-glosa dados pelo alinhamento e na correspondência entre as classes gramaticais do lado português e as do lado da LGP. Essa correspondência é marcada por um número, chamado de **número de correspondência**. As classes gramaticais dos lemas que compõem os gestos

compostos terão o mesmo número para preservar o alinhamento. Assim, as classes gramaticais dos lemas *haver* e *grande* terão o mesmo número de correspondência que a classe gramatical do gesto composto *TER-MUITO*. A regra morfossintática para o predicado do par de frases 18 e 19 está representada em 22. No lado português dessa regra, as classes gramaticais V1 e ADJ1 são as classes dos lemas *haver* e *grande* e, no lado da LGP, V1 é a classe do gesto *TER-MUITO*. A regra indica ainda uma troca do lema *desenvolvimento* representado pela etiqueta *N2* para o fim da frase em LGP.

(22) V1 ADJ1 N2 ADJ4 ADJ3 ADJ5 ADJ6 $\rightarrow$ V1 N4 N3 N5 N6 N2

Ao todo foram construídas 66 regras morfossintáticas, sendo que 18 são relativas a sujeitos, 46 de predicados e 2 de modificadores de frase, e 39 regras frásicas, 5 associadas a frases interrogativas, 3 de frases negativas e 31 de frases declarativas afirmativas.

As regras frásicas são guardadas em ficheiros *CSV* diferentes para cada tipo de frase, enquanto que as regras morfossintáticas são divididas em 3 ficheiros, respeitantes a cada elemento frásico. A primeira coluna desses ficheiros contém o lado português da regra e a segunda coluna o lado da LGP. Os ficheiros com as regras morfossintáticas apresentam uma terceira coluna que identifica o tipo de frase associado à regra. Ao guardar as regras em ficheiros, o tradutor não tem de gerar as regras de novo para poder traduzir.

5.1.6 Estatísticas das regras automáticas

Durante a construção automática das regras de tradução, procedeu-se à contagem da ocorrência de cada regra, para cada tipo de frase.

Relativamente às regras dos constituintes morfossintáticos, duas regras são consideradas iguais se o tipo de frase das duas for o mesmo e se as sequências dos seus constituintes forem exactamente iguais, assim como as correspondências entre eles (assinaladas pelos números). Por exemplo, as regras 23 e 24 não são a mesma porque a correspondência entre os constituintes não é a mesma. No lado português da primeira regra, *V1* corresponde a *V1* no lado da LGP, *ADJ1* corresponde também a *V1*, contudo, o adjetivo (ADJ2) na segunda regra não corresponde a *V1* mas sim a *N2*, quebrando a igualdade das regras.

(23) V1 ADJ1 N2 $\rightarrow$ V1 N2

(24) V1 ADJ2 N2 $\rightarrow$ V1 N2

Por outro lado, as regras 25 e 26 são a mesma. Apesar de os números não serem os mesmos, o mapeamento entre os constituintes é respeitado, ou seja, os verbos *V3* e *V2* alinham-se com os nomes *N3* e *N2*, os adjetivos *ADJ2* e *ADJ3* correspondem a *ADJ2* e *ADJ3*, etc.

(25) V3 N1 ADJ2 N3 $\rightarrow$ N3 ADJ2 N1

(26) V2 N1 ADJ3 N2 → N2 ADJ3 N1

Quanto à contagem das regras dos elementos frásicos, duas regras são iguais se o tipo da frase for o mesmo e a ordem frásica na regra for igual. Por exemplo, as regras em 27 e em 28, apesar de representarem a mesma ordem frásica, correspondem a tipos de frases diferentes e por isso, contabilizam-se em separado, como regras diferentes.

(27) SVO → SVO, para frases declarativas negativas.

(28) SVO → SVO, para frases declarativas afirmativas.

As estatísticas das regras com a ordem dos constituintes são guardadas em 3 ficheiros JSON, consoante se as regras pertencem a sujeito, predicado ou modificador de frase. Ao passo que as da estrutura frásica estão representadas num único ficheiro JSON. Estas estatísticas serão usadas no módulo de tradução, descrito na Secção seguinte (5.2). Além da sua importância no tradutor, apresentam informações linguísticas relevantes para o estudo de alguns fenómenos gramaticais da LGP, como a ordem frásica canónica[12] ou base.

5.1.7 Dicionário bilingue de português e língua gestual portuguesa

Do alinhamento das palavras com os gestos do corpus (Secção 2.3) desenvolveu-se automaticamente um dicionário bilingue de português e LGP para auxiliar a transferência lexical no tradutor (Secção 5.2.5.A), i.e., o mapeamento entre o léxico português e o léxico da LGP. No total foram alinhados 163 pares palavra-gesto, a maioria corresponde a pares palavra-glosa (*arte* e *ARTE*) mas existem ainda pares semanticamente relacionados como, *religião* e *IGREJA* e ainda pares com gestos compostos como *muita riqueza* e *MUITO-RICO*. Dado que o alinhamento não é perfeito, o dicionário foi posteriormente revisto com base nas informações do vídeo transcritas. Após a revisão e eliminação de correspondências erróneas como *século* e *ARTE*, o dicionário apresenta 102 entradas. Contudo, à medida que for processada informação nova, extraída do ELAN, serão adicionadas novas entradas no dicionário, pelo que este deverá ser sempre revisto.

Dado que a derivação dos gestos em LGP difere da derivação das palavras na língua portuguesa, a um mesmo gesto poderão estar associados várias flexões de uma palavra, por exemplo, a diferentes flexões de tempo verbais está associado o mesmo gesto, como *tive* e *tenho* cujo gesto correspondente a ambas é *TER*. Assim, para tornar o mapeamento entre os léxicos de português e de LGP mais flexível no tradutor, o dicionário contém, além da palavra e o respetivo gesto, o lema da palavra em português (Anexo D).

[12] A ordem canónica estabelece-se a partir de frases declarativas e afirmativas, que não tenham sofrido topicalização.

5.1.8 Regras manuais

Um conjunto de regras manuais complementa as regras automáticas anteriormente descritas. Com base nas características gramaticais da LGP listadas na Secção 2 construíram-se 16 regras manuais que garantem que a ordem de constituintes com determinadas subclasses esteja de acordo com a ordem dos constituintes da LGP (**regras manuais sintáticas**) e integram particularidades da língua relacionadas com a morfologia das palavras (**regras manuais morfológicas**), como a marcação do género feminino, dos tempos verbais e do grau do substantivo e como as expressões faciais gramaticais relativas às frases negativas e interrogativas. Os processamentos aqui descritos foram realizados com base nas informações morfológicas das palavras, mais precisamente em informações de flexão nominal em grau, género, número, flexão verbal em modo e tempo e nas classes e subclasses das palavras identificadas na fase de análise (Secção 5.2.4).

5.1.8.A Regras manuais sintáticas

Três das regras implementadas consistem na ordenação de determinantes possessivos e de elementos que formam o plural como numerais e advérbios de quantidade. Em LGP, estes são produzidos depois do substantivo. A ferramenta de classificação morfossintática, Freeling, usada na fase de análise (Secção 5.2.4) só identifica dois tipos de advérbios (advérbios de negação e gerais), pelo que os advérbios de quantidade (muito, mais, bastante, etc.) foram previamente identificados numa lista.

Para o tratamento de pronomes e advérbios interrogativos em interrogativas parciais[13] foram seguidos os resultados de um estudo preliminar sobre o comportamento sintático destes tipos de interrogativas na LGP [62], devido à falta de uma gramática e estudos recentes. Este estudo foi realizado com base em 30 perguntas licitadas a 5 participantes (todos surdos e gestuantes nativos de LGP). Apesar de ser preliminar verifica-se uma tendência em marcar as interrogativas parciais de objeto[14] colocando o elemento QU- no fim da frase simultaneamente à expressão facial como no exemplo 29 retirado do estudo. Contudo, os resultados das estruturas de interrogativas parciais de sujeito[15] são inconclusivos, variam principalmente entre quatro estruturas. Seriam necessários mais dados para se poder verificar um eventual padrão na marcação de deste tipo de interrogativas parciais. Devido à complexidade das outras estruturas, decidiu-se seguir e implementar no tradutor a estrutura em 30. Nos seguintes exemplos, a expressão facial é identificada pela linha e o *q* por cima das glosas.

(29) GATO MORDER $\overline{QUEM?}^{q}$ (Quem é que o gato está a morder?)

(30) MORDER VACA $\overline{QUEM?}^{q}$ (Quem está a morder a vaca?)

[13] De acordo com a página ciberdúvidas interrogativas parciais possuem um elemento interrogativo (QU-), no qual, fazem parte os pronomes e advérbios interrogativos, como Quem e Onde.
[14] Nas interrogativas parciais de objeto, a questão recai sobre o objeto.
[15] Nas interrogativas parciais de sujeito, a questão recai sobre o sujeito.

Assim, sempre que forem detetados pronomes interrogativos e advérbios interrogativos, estes são colocados no fim da sequência de glosas acompanhados pela expressão facial. Dependendo se o elemento QU- recai sobre o sujeito ou o objeto existe uma alteração da ordem sintática do sujeito ou objeto e verbo. No caso das interrogativas parciais de sujeito, o verbo encontra-se antes do objeto (ver exemplo 30), enquanto que, nas interrogativas parciais de objeto, o verbo segue o sujeito como no exemplo 29. Nas interrogativas globais, a ordem dos constituintes é ditada pelas regras extraídas do corpus. Para se proceder ao tratamento destas interrogativas, primeiro distinguiu-se interrogativas parciais de interrogativas globais através de uma regra simples: se houver um pronome interrogativo ou advérbio interrogativo, então é uma interrogativa parcial. Depois, distinguiu-se interrogativas parciais de sujeito das interrogativas parciais de objeto, verificando o elemento sintático (sujeito ou objeto) presente na frase. Se possuir sujeito, então estamos perante uma interrogativa parcial de objeto, caso contrário, é uma interrogativa parcial de sujeito.

Enquanto que na marcação das interrogativas parciais, advérbios e pronomes interrogativos são colocados no final da sequência de glosas, os advérbios de tempo como *Ontem* são produzidos primeiro em LGP. Tal como para os advérbios anteriores, criaram-se 3 listas para identificar advérbios de tempo com conotação do passado (ontem, antes, etc.), do futuro (amanhã, etc.) e com conotação do presente (hoje, agora, etc.). Contudo, esta abordagem é limitada, expressões temporais como *No ano passado* não são identificadas pelo tradutor.

A geração da negação e das interrogativas parciais estão associadas a regras manuais mais limitativas devido à irregularidade da sua marcação (ver Secção 2.2.11). Na pesquisa realizada para esta dissertação não estão presentes estudos que indiquem em que contexto se recorre a cada uma das opções de marcação da negação, pelo que decidiu-se optar pela adição do marcador não manual *headshake* em simultâneo à componente manual *NÃO*, por ser o marcador manual mais frequente na LGP [15], como está exemplificado em 31.

(31) AMANHÃ DT(C-A-R-O-L-I-N-A) VESTIR $\overline{NAO}^{headshake}$ (Amanhã, a Carolina não se vai vestir.)

Dado que os verbos *Ser* e *Estar* não são produzidos em LGP, estes são removidos da tradução. As preposições são igualmente eliminadas por não serem representadas em LGP isoladamente. Ainda remove-se a conjunção coordenativa copulativa *e*, por ser uma conexão prosódica, expressa não manualmente pela expressão facial neutra.

5.1.8.B Regras manuais morfológicas

Os nomes próprios são produzidos com recurso à datilologia, ou seja, ao alfabeto manual. A identificação deste fenómeno é baseada na notação usada no corpus (ver Tabela 2.2), i.e., nomes próprios são representados usando a notação *DT(N-O-M-E)*. Assim, Maria representa-se como *DT(M-A-R-I-A)* na

sequência de glosas. A ferramenta Freeling identifica substantivos próprios compostos (Secção 4) como *Maria Silva*, assim, foi necessário criar uma notação para identificá-los. Decidiu-se seguir a representação de nomes compostos usada pela ferramenta, na qual, um _ é usado para separar os constituintes do nome próprio, resultando na seguinte estrutura: *DT(M-A-R-I-A-_-S-I-L-V-A)*. Antes desta representação, explorou-se esta: *DT(M-A-R-I-A) DT(S-I-L-V-A)*, mas tornava impossível a distinção destes casos (de nomes compostos) dos casos em que várias pessoas são mencionadas numa mesma frase, como em *A Maria e o Silva ...*, cuja tradução seria, também, *DT(M-A-R-I-A) DT(S-I-L-V-A)...*.

De acordo com a Secção 2.2.3, a marcação do feminino na LGP realiza-se através da prefixação do gesto *MULHER* ao gesto base, excepto em alguns substantivos, que possuem gestos diferentes para cada género, como *GALO* e *GALINHA* e *ENFERMEIRO* e *ENFERMEIRA*. Tanto em português como na LGP, os substantivos no feminino ou derivam da forma masculina ou têm um lema próprio (*GALO* e *GALINHA*) e, para simplificar, assumiu-se que se uma palavra feminina em português tem o seu próprio lema, também tem o seu próprio gesto na LGP (e se deriva da forma masculina em português, na LGP é formado por composição com o gesto *MULHER*). Na LGP, a marcação do género ocorre apenas em substantivos que designam seres animados. Idealmente, a marcação do género deveria ser apoiada por um dicionário bilingue completo mas, na sua ausência, recorreu-se a uma heurística simples para distinguir substantivos animados de inanimados baseada no lema da palavra no feminino. Caso a palavra no feminino seja igual ao seu lema, então a palavra poderá corresponder a um ser inanimado como mesa ou poderá corresponder a um ser animado mas com lemas diferentes para cada género como vaca e boi. Caso contrário, é um ser animado possuindo sexo, como irmã, cujo lema é irmão. No primeiro caso, em que a palavra é igual ao seu lema, as palavras foram convertidas em glosas sem qualquer alteração. No segundo, em que a palavra é diferente do seu lema, acrescentou-se o gesto *MULHER* ao gesto base, por exemplo, a palavra irmã corresponde à sequência de gestos *MULHER* e *IRMÃO*. Contudo, existem exceções como a palavra égua que, apesar de ser igual ao seu lema, na LGP corresponde à composição dos gestos *MULHER + CAVALO* e, por outro lado, a palavra enfermeira que, apesar de ter o lema enfermeiro, não sofre marcação por prefixação: os gestos *ENFERMEIRO* e *ENFERMEIRA* são gestos diferentes. Estas exceções foram identificadas num ficheiro de texto, de forma a permitir futura extensão desta lista. Os substantivos foram tratados como masculinos nos casos em que o seu género em português era ambíguo e quando a ferramenta foi incapaz de o determinar.

Dado que a ferramenta identifica os casos de grau do tamanho dos substantivos, o tratamento dos mesmos foi facilitado consistindo no processamento da palavra conforme a etiqueta morfossintática. O grau dos substantivos é identificado na última letra da etiqueta morfossintática, caso essa letra seja um *A* então estamos perante o caso aumentativo, caso seja um *D* estamos perante o caso diminutivo. Posteriormente ao lema da palavra são adicionados os gestos *GRANDE* ou *PEQUENO*, conforme o grau do tamanho do substantivo. No caso de uma palavra estar no feminino e com grau de tamanho,

a glosa *MULHER* é ordenada para antes do lema e do gesto que marca o tamanho. Por exemplo, *LEOAZINHA* corresponde à sequência de glosas *MULHER LEÃO PEQUENO*.

Conforme [12] e [63], a marcação do passado e futuro é realizada através de informações temporais, quando estas estão presentes nas frases, explicitadas através de advérbios de tempo (ontem) ou expressões temporais (na semana passada). Caso a frase não contenha informação temporal além da que está presente no tempo verbal, adicionam-se os gestos *PASSADO* ou *FUTURO* consoante o mesmo. Estas informações temporais (advérbios e expressões adverbiais de tempo e os gestos *PASSADO* e *FUTURO*) são colocadas no inicio da frase em LGP. Mais uma vez, porque a análise morfossintática realizada pela ferramenta Freeling não pormenoriza o tipo de advérbios, a identificação de advérbios e expressões temporais sustenta-se através de duas listas com informações temporais referentes ao passado (antigamente, antes, hoje de manhã) e informações temporais referentes ao futuro (logo, amanhã, doravante,em breve). Esta solução não cobre todos os possíveis advérbios e expressões temporais da língua portuguesa. Assim, caso a frase em português contenha um advérbio ou expressão facial que não esteja presente nessas listas, então a sequência de glosas será composta por um dos gestos PASSADO ou FUTURO, além do advérbio ou expressão temporal presente na frase em português. Por exemplo, dada a frase *Na próxima segunda-feira irei telefonar à minha mãe.* e a inexistência na lista da expressão temporal *Na próxima segunda-feira*, o resultado do tradutor corresponde a *FUTURO PRÓXIMA SEGUNDA-FEIRA TELEFONAR MÃE MEU*.

As expressões faciais gramaticais podem persistir ao longo da frase toda, como acontece no caso das interrogativas globais [1], podem acompanhar certos gestos ou podem aparecer simultâneamente a um gesto. Havendo esta variabilidade na duração das expressões faciais criou-se uma notação para identificar o começo e o fim da expressão facial, que gestos engloba e qual a expressão facial.Para identificar a duração da expressão facial usou-se chavetas, na qual, a chaveta aberta indicia o inicio da expressão facial e a chaveta fechada o fim da mesma. Por sua vez, a expressão facial aparece entre parênteses curvos após a identificação do término da expressão facial. Por exemplo, o resultado do tradutor para a frase identificada em 29 corresponde a *GATO MORDER* {*QUEM*}*(q)*, em que q corresponde ao gesto não manual da interrogativa: levantar o queixo, inclinar a cabeça para trás e franzir as sobrancelhas. O mesmo se aplica na representação da expressão facial *headshake* nas frases negativas. Por exemplo, a frase em 31 seria representada como *AMANHÃ DT(C-A-R-O-L-I-N-A) VESTIR* {*NÃO*}*(headshake)*.

5.2 Módulo de tradução automática

O tradutor é composto por quatro fases principais identificadas na Figura 5.6. O pré-processamento consiste na segmentação do *input* em frases e no tratamento de expressões que não deverão ser

processadas pelo tradutor para que a ordem das suas unidades lexicais e o seu significado não sejam alterados (Secção 5.2.3). De seguida, a frase de entrada é analisada sintática e morfologicamente (Secção 5.2.4). Na fase de transferência gramatical (Secção 5.2.5), primeiro o léxico português é mapeado para o léxico da LGP (transferência lexical), de seguida as regras manuais sintáticas (Secção 5.1.8.A) e as regras construídas a partir do corpus de referência da Universidade Católica (Secção 5.1) são aplicadas à frase de entrada, convertendo a estrutura da frase em português na sua estrutura em LGP (transferência sintática). Na última fase (Secção 5.2.6), são aplicadas as regras manuais relacionadas com a morfologia das palavras e com a ordem sintática de constituintes (Secção 5.1.8.B). Marcação das expressões faciais, feminino, tempos verbais e do grau do substantivo são exemplos dessas regras. O resultado do tradutor é uma sequência de gestos em glosas, que representa a frase em LGP.

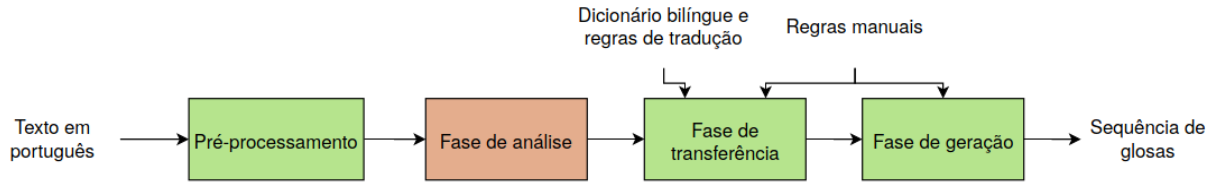


Figura 5.6: Arquitetura do tradutor.

Nas próximas secções, os procedimentos de cada uma destas componentes serão detalhados e exemplificados através da frase exemplo em 32.

(32) A Diana perdeu o seu gatinho ontem.

5.2.1 Estrutura do *input*

O tradutor está preparado para receber uma ou mais frases em português, no entanto, só processa uma de cada vez. Caso o *input* seja um conjunto de frases, então, este é segmentado em frases individuais na componente *pré-processamento*, que serão posteriormente traduzidas individualmente. A frase de partida em 32 é um exemplo de um *input*.

5.2.2 Estrutura da sequência de glosas (*output*)

O resultado do tradutor consiste numa sequência de gestos em glosas com marcadores que identificam expressões faciais, gestos compostos e datilologia (no caso da soletração de nomes próprios). A representação dos gestos compostos (por mais do que uma palavra portuguesa) e datilologia seguem as convenções definidas para anotação desses elementos no corpus de referência. Na Tabela 2.2, descrevem-se algumas dessas convenções, por exemplo, os gestos compostos são identificados

por um hífen, como *POR-FAVOR*. O uso das convenções do corpus permite que haja uma compatibilidade entre as informações que saem do corpus e as informações do tradutor, evitando conversões intermédias. A notação usada para representar as expressões faciais está descrita na Secção 5.1.8.B.

Dada a frase exemplo em 32, o resultado do sistema é *ONTEM DT(D-I-A-N-A) PERDER GATO PEQUENO SEU*, em que DT() indica a "soletração" do nome próprio *Diana*.

5.2.3 Pré-processamento do input

Dois pré-processamentos são aplicados ao input antes de este passar para as fases seguintes. Primeiro, se o input consistir em várias frases, este é segmentado em frases e cada uma é tratada individualmente pelas próximas fases. Contudo, existem expressões cuja ordem das palavras não deverá ser alterada para que o seu significado também não o seja. Para tal, estas expressões são pré-processadas. Exemplos destas expressões são *bom dia* e *por favor*. Como na LGP as preposições não são produzidas individualmente, a preposição *por* em *por favor* seria removida e o resultado da tradução seria, erroneamente, *favor*. Isto acontece porque as excepções são compostas por unidades lexicais, que são tratadas pelo tradutor individualmente. Para resolver esta situação, estas excepções previamente discriminadas numa lista, são convertidas para uma notação em que o tradutor as veja como uma só unidade lexical. Seguindo as convenções do corpus para gestos compostos por mais do que uma palavra em português, as unidades lexicais das excepções são ligadas por um hífen, ou seja, *por favor* é representado por *por-favor*.

A frase em 32 passa para a fase seguinte sem sofrer qualquer dos processamentos, sendo o resultado desta fase: *A Diana perdeu o seu gatinho ontem.*

5.2.4 Fase de análise

Os procedimentos realizados nesta fase são comuns ao módulo de construção de regras de tradução automáticas (Secção 5.1.3), pelo que não serão novamente detalhados aqui. A transferência entre a estrutura sintática da frase em português e a estrutura em LGP realiza-se por aplicação das regras definidas no módulo de construção de regras de tradução na Secção 5.1. A cada **elemento frásico** (sujeito, predicado e modificador da frase) estão associadas regras automáticas. Por forma a poderem ser aplicadas, a frase em português é analisada sintática e morfologicamente com recurso às mesmas ferramentas de processamento de texto usadas no módulo anterior (Secção 5.1.3.A). Na análise morfossintática identificam-se as classes e subclasses gramaticais, assim como aspetos de flexão das palavras da frase, enquanto que, na análise sintática a frase é dividida nos seus elementos frásicos.

5.2.4.A Análise morfossintática

Nesta etapa a frase de entrada é analisada morfossintaticamente pela ferramenta Freeling, ficando-se a conhecer as classes e subclasses gramaticais (pronomes possessivos, determinantes demonstrativos, etc.) e, ainda aspetos de flexão (flexão em número, género, tempo verbal, etc.) de cada palavra.

A análise morfossintática da frase exemplo em 32, *A Diana perdeu o seu gatinho ontem.*, é *DA0FS0 NP00000 VMIS3S0 DA0MS0 DP3MSS NCMS00D RG Fp*. O significado destas etiquetas pode ser consultado no manual de utilizador da ferramenta.

5.2.4.B Análise sintática

Nesta etapa, a frase é segmentada nos seus elementos frásicos (sujeito, predicado e modificador da frase). Este procedimento encontra-se detalhado na Secção 5.1.3.C.

Assim, a frase *A Diana perdeu o seu gatinho ontem.*, com a ordem SVO, é representada pelos seus elementos frásicos: sujeito, *a Diana* e predicado, *perdeu o seu gatinho ontem*.

5.2.4.C Pós-processamento

No final desta fase removem-se os determinantes artigos (definidos e indefinidos), preposições e sinais de pontuação dos resultados das etapas anteriores e convertem-se as etiquetas resultantes das análises morfossintática e sintática para as do corpus, uniformizando-as com as usadas nas regras automáticas. Antes de a pontuação ser removida, o tipo de frase (declarativa afirmativa, negativa, exclamativa ou interrogativa) é determinado e guardado por ser necessário na transferência sintática (Secção 5.2.5.B). Assim, no final da fase de análise, a frase exemplo é representada por *N V DET N ADV* (*Diana perdeu seu gatinho ontem*), sendo o sujeito composto por *N* (*Diana*) e o predicado por *V DET N ADV* (*perdeu seu gatinho ontem*).

Desta forma, reunem-se as condições necessárias para se iniciar a fase de transferência gramatical.

5.2.5 Fase de transferência

Esta fase divide-se em duas etapas, nas transferências lexical e sintática. Estas etapas realizam-se com base nas informações gramaticais extraídas do corpus (Secção 5.1). Na primeira, o léxico do português é convertido no léxico da LGP com base no dicionário bilingue descrito na Secção 5.1.7, enquanto que na transferência sintática, as regras de tradução automáticas (Secção 5.1.5) e as regras manuais sintáticas (Secção 5.1.8.A) são aplicadas à frase em português, de forma a converter a sua estrutura na estrutura da LGP.

5.2.5.A Transferência lexical

O mapeamento entre o léxico português e o léxico da LGP realiza-se através do dicionário bilingue de português e LGP. Tal como referido na Secção 5.1.7, as entradas deste dicionário correspondem a palavras e lemas do léxico português, bem como os respetivos gestos. A transferência lexical é então, baseada na pesquisa de porções da frase em português que poderão ser simplesmente lemas de unidades lexicais, ou um conjunto de lemas, como *haver grande*. Caso o lema ou sequência de lemas esteja no dicionário, então será substituída pelo gesto correspondente, caso contrário, será convertida em glosa na fase de geração. Esta pesquisa por lema, ao invés pela palavra em si, expande a cobertura do léxico possível de ser traduzido pelo dicionário. Desta forma, quer a expressão *houve um grande*, quer a expressão *há um grande* serão transformadas em *TER-MUITO*, dado que o lema de *houve* e *há* é *haver* e que existe a sequência de lemas *haver grande* no dicionário. Contudo, existe uma diferença entre ambas que se centra no seu tempo verbal e esta diferença não deverá ser esquecida para que a tradução não seja incorreta. Para tal, nesta fase preservam-se as informações morfológicas (nomeadamente, o tempo verbal) das palavras em português, para posteriormente serem devidamente processadas na fase de geração (Secção 5.2.6). Concretizando, na fase de geração seria acrescentado à primeira expressão o gesto *PASSADO* para marcar o tempo verbal passado, ficando *PASSADO TER-MUITO*, que difere da segunda expressão traduzida como *TER-MUITO*, por estar no presente.

Esta abordagem baseada nos lemas não se aplica nos casos em que a palavra é um substantivo comum. Dado que se está a trabalhar com glosas, as glosas *ÁRVORE* e *ÁRVORES* correspondem a gestos diferentes (ver Secção 2.2.5), pelo que a pesquisa no dicionário para estes casos realiza-se através da palavra em português e não do seu lema, caso contrário, a palavra *árvores* seria traduzida erroneamente como ÁRVORE. Esta pesquisa lexical não tem em conta a semântica da palavra portuguesa no dado contexto, provocando erros na tradução lexical. Por exemplo, para o tradutor, o verbo *chamar* tem o mesmo significado nas frases *Como te chamas?* e *Chamas um táxi por mim, por favor?*, levando a que o resultado da tradução para esse verbo, em ambas as frases, seja *NOME*, conforme o dicionário.

A frase de partida não sofre alterações nesta fase.

5.2.5.B Transferência sintática

Nesta fase as classes gramaticais e os constituintes da frase (sujeito, verbo e objeto) são ordenados de acordo com as regras manuais sintáticas e as regras automáticas, consoante o tipo da frase (declarativas afirmativas (CAN), negativas (NEG), interrogativas (INT) e exclamativas (EXCL)).

É importante clarificar que as operações desta fase não se realizam sobre a frase em português mas sobre os seus elementos frásicos (sujeito, predicado e modificador de frase), divididos na análise sintática (Secção 5.2.4.B). Assim, o que é recebido nesta fase são as **estruturas sintáticas** de cada

elemento frásico, exemplificadas em 33 para o sujeito e em 34 para o predicado da frase exemplo em 32, *A Diana perdeu o seu gatinho ontem*.

(33) Estrutura sintática do sujeito: N (N é Nome)

(34) Estrutura sintática do predicado: V DET N ADV (V é Verbo, DET é Determinante, N é Nome e ADV é Advérbio)

Esta segmentação da frase por funções sintáticas ocorre porque as regras de tradução automáticas dizem respeito a esses três elementos frásicos.

Tal como indicado na Secção 5.1.5, as regras são compostas por dois lados, o lado origem, chamado de **lado português** e o lado destino, o **lado da LGP**. Por exemplo, para a regra *V1 ADJ3 PRO2* → *V1 PRO2 ADJ3*, o lado português corresponde ao lado esquerdo da seta e o lado da LGP ao lado direito. Os números marcam a correspondência entre as classes gramaticais dos dois lados, chamados de **números de correspondências**.

A transferência da estrutura sintática da frase em LP para a estrutura em LGP realiza-se, então, em três passos, como indica o esquema na Figura 5.7. As classes gramaticais de cada elemento frásico são ordenadas com base nas regras manuais sintáticas (Secção 5.1.8.A) e na regra de tradução morfossintática que melhor se ajusta à estrutura sintática do elemento. Os elementos frásicos são ordenados conforme a ordem frásica da LGP mais frequente no corpus, dependendo do tipo da frase.

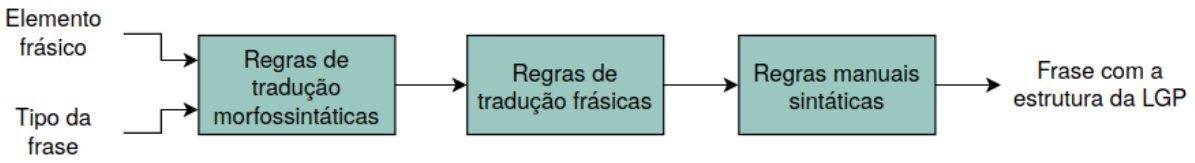


Figura 5.7: Passos da transferência sintática.

A regra de tradução que melhor se ajusta ao elemento frásico corresponde àquela cuja estrutura sintática do lado português se assemelha mais à estrutura do elemento frásico da frase em português, a partir de agora tratado como *frase* para simplificar. Assim, para cada regra de tradução calcula-se a semelhança entre o lado português e a estrutura sintática da frase com o algoritmo modificado da Distância de Edição (ou Distância de Levenshtein [64]). Este procedimento garante que a todas as frases de entrada seja atribuída uma regra de tradução morfossintática.

Antes de se proceder ao cálculo das distâncias, tanto a estrutura da frase como a das regras do lado da língua portuguesa são convertidas para o seguinte formato: *CL1 CL2 CL3 Tipo_da_frase*, em que *CL* são classes gramaticais e *Tipo_da_frase* corresponde a uma das seguintes hipóteses: exclamativa (EXCL), declarativa afirmativa (CAN), declarativa negativa (NEG) e interrogativa (INT). Desta forma, as **estrutura do sujeito e do predicado** da frase exemplo são convertidas para:

(35) Sujeito: N CAN

(36) Predicado: V DET N ADV CAN

Tendo ambas as estruturas uniformizadas, o passo seguinte consiste no cálculo da Distância de Edição entre todas as regras e a frase. No final, a regra com **menor distância** é a mais semelhante à frase.

A Distância de Edição é uma medida de semelhança entre duas sequências[16], que permite saber que operações devem ser feitas para que as duas sequências fiquem iguais. As operações possíveis são inserção, remoção e substituição. Os custos implementados para estas operações são de 1, excepto no caso é que o tipo de frase é substituído, em que o custo é 2. O objetivo é igualar a regra do lado português à frase em português, pelo que as operações são aplicadas à regra de tradução.

Após calcular a Distância de Edição entre a frase e cada regra, é escolhida a regra com menor distância, pois essa é a mais semelhante à frase. Em caso de empate, seguem-se os seguintes critérios por ordem:

1. Escolhe-se a regra mais frequente no corpus com base nas estatísticas recolhidas;
2. Escolhe-se a maior regra;
3. Escolhe-se a regra que vem primeiro alfabeticamente.

Estes critérios de desempate são arbitrários, mas garantem que a escolha da regra seja consistente.

As regras automáticas escolhidas para cada elemento frásico do exemplo são:

(37) Sujeito: N1 CAN $\rightarrow$ N1 CAN

(38) Predicado: V1 N2 ADV3 CAN $\rightarrow$ V1 ADJ3 N2 CAN

As distâncias para as regras anteriores (37 e 38) são de 0 e 1, respetivamente.

Tendo a regra que melhor se ajusta à frase, o próximo passo consiste na aplicação das operações dadas pela Distância de Edição, à regra de tradução para torná-la igual à frase. Relembrando que as regras automáticas são compostas por dois lados (português e LGP), as operações aplicadas a um lado são também aplicadas ao outro.

Para o exemplo, apenas a regra de tradução do predicado sofre alterações, pois a regra do lado português do sujeito em 37 (N CAN) e a estrutura do sujeito em 35 da frase (N CAN) são iguais. Para igualar a regra de tradução do predicado em 38 à estrutura do predicado em 36 é necessário inserir um DET (determinante) depois do V1 no lado português da regra. Contudo, ao inserir nesse lado, também deverá ser inserido no lado da LGP. Para realizar inserções no lado da LGP decidiu-se seguir uma heurística simples: o elemento a adicionar no lado da LGP é inserido a seguir à classe gramatical com

[16] Explicação detalhada do algoritmo: web.stanford.edu/class/cs124/lec/med.pdf

o número de correspondência igual à classe gramatical anterior ao valor inserido no lado português. Assim, insere-se o DET depois de V1 tanto no lado português da regra como no lado LGP da regra. Para marcar a correspondência adiciona-se ao DET o número 4. O resultado final das operações é:

(39) Sujeito: N1 CAN → N1 CAN

(40) Predicado: V1 DET4 N2 ADV3 CAN → V1 DET4 ADJ3 N2 CAN

As duas restantes operações são mais simples de realizar, o constituinte a remover ou a substituir no lado LGP da regra é aquele com o mesmo número de correspondência do constituinte que foi removido/substituído no lado português.

Existe ainda um passo a realizar antes de se aplicar a regra, que é uniformizar os dois lados da regra. A regra de tradução em 40 diz que o advérbio (ADV) do lado português deverá ser convertido no seu adjetivo (ADJ), contudo não existem ferramentas de processamento de português e conhecimento sobre estes fenómenos que permitam esta conversão e, por isso, estes fenómenos linguísticos são ignorados, ou seja, o advérbio continuará a ser advérbio. Este é o resultado depois da uniformização da regra:

(41) Sujeito: N1 CAN → N1 CAN

(42) Predicado: V1 DET4 N2 ADV3 CAN → V1 DET4 ADV3 N2 CAN

Por fim, aplicam-se as regras, fazendo corresponder cada constituinte morfossintático à palavra. Por exemplo, para o predicado, V1 corresponde a *perdeu*, DET4 a *seu*, N2 a *gatinho* e ADV3 *ontem*. No caso do sujeito, N1 corresponde a *Diana*. A regra de tradução do predicado dita uma alteração na ordem do advérbio (ADV) e do substantivo comum (N). Assim o resultado final da ordenação dos constituintes morfossintáticos equivale a:

(43) Sujeito: Diana

(44) Predicado: perdeu seu ontem gatinho

Os elementos frásicos, com uma nova estrutura, são unidos para formarem a frase em LGP. Esta união realiza-se com base na ordem frásica mais frequente no corpus de acordo com o tipo da frase, ou seja, se a frase em português for declarativa negativa e a ordem frásica mais frequente para esse tipo de frase for SOV, então os elementos sintáticos são ordenados por essa ordem, primeiro sujeito, seguido de objeto e, no fim, o verbo. As frequências das diferentes ordens frásicas para o mesmo tipo de frase (declarativas afirmativas (CAN), declarativas negativas (NEG), interrogativas (INT) e exclamativas (EXCL)) no corpus são um parâmetro de entrada desta fase.

Dado que para as frases declarativas afirmativas contabilizaram-se 7 estruturas SVO e 2 SOV, a tradução para LGP da frase *A Diana perdeu o seu gatinho ontem* terá a ordem SVO. Tendo em conta a composição dos seus elementos frásicos em 43 (*Diana*) e 44 (*perdeu seu ontem gatinho*), o resultado

da aplicação da estrutura mais frequente do corpus é *Diana perdeu seu ontem gatinho*. Contudo, e seguindo a premissa de estudos anteriores, em que a estrutura frásica base da LGP é SOV, adicionou-se uma opção de escolha entre a estrutura mais frequente do corpus ou a estrutura SOV. Se fosse escolhida esta estrutura, então o resultado da transferência sintática seria *Diana seu ontem gatinho perdeu*.

Por último, aplicam-se as regras manuais sintáticas (Secção 5.1.8.A). Dado que em LGP os advérbios de tempo são produzidos no início e os determinantes possessivos procedem o substantivo, o resultado da transferência gramatical da frase em 32 é *Ontem Diana perdeu gatinho seu*.

5.2.6 Fase de geração

Aqui, o léxico é convertido em glosas e são aplicadas as regras manuais morfológicas (Secção 5.1.8.B) que integram particularidades da língua relacionadas com a morfologia das palavras, como a marcação do género feminino, graus do tamanho do substantivo, datilologia de nomes próprios, marcação dos tempos verbais e expressões faciais gramaticais (das frases interrogativas e negativas). Desta fase sai uma sequência de glosas com marcadores adicionais que identificam expressões faciais e palavras soletradas. Assim, a tradução da frase em 32 é *ONTEM DT(D-I-A-N-A) PERDER GATO PEQUENO SEU*, em que a notação *DT()* indica que o nome próprio Diana é "soletrado".

6

Avaliação experimental

Conteúdo

Os procedimentos experimentais e os resultados da avaliação do desempenho do presente sistema de tradução automática são descritos neste capítulo. Dado que o sistema se divide em dois módulos, primeiro descreve-se a experiência e os resultados para o módulo de regras de tradução, denominada como **experiência 1** (Secção 6.3) e depois a experiência e os resultados para o módulo de tradução, apelidada como **experiência 2** (Secção 6.4). O primeiro módulo é avaliado comparando manualmente as traduções produzidas pelo sistema com as traduções do corpus de referência. Para avaliar a qualidade da tradução do sistema proposto (sistema PE2LGP) conduziram-se duas avaliações, uma automática comparando o output do sistema com uma tradução que se sabe ser correta e outra manual, com base na opinião de peritos. Antes descrevem-se os corpora usados na avaliação do sistema (Secção 6.1) e apresentam-se as medidas de avaliação, na Secção 6.2.

6.1 Corpora

6.1.1 Corpus de desenvolvimento

Simultaneamente ao desenvolvimento do tradutor, foi criado um corpus de desenvolvimento, que acabou por ser composto por 75 frases em português (Anexo E). Este corpus não foi usado apenas para testar aspetos técnicos mas também para, durante o desenvolvimento do tradutor, se ter uma ideia básica do seu desempenho em cada passo do processo de tradução.

6.1.2 Corpus de teste

O corpus de teste (ou coleção dourada) foi criado por uma intérprete de português e LGP. É composto por 58 frases simples em português e as correspondentes traduções em LGP. O corpus é de domínio aberto. Para algumas frases em português foi anotada mais do que uma tradução possível, mas não se procurou obter todas as traduções possíveis.

Apesar de ser constituído por 58 frases, estas têm origem em 19 frases declarativas, portanto, as restantes 39 frases correspondem às formas negativas e interrogativas das 19 frases originais.

Por fim, antes de se proceder à experiência 2, as sequências de glosas nas traduções do corpus foram adaptadas para a estrutura do output do tradutor, seguindo as convenções descritas na Secção 5.2.2. Esta etapa foi realizada com o acompanhamento da intérprete para confirmar que as frases convertidas para a notação do output do sistema continuavam a corresponder à frase anotada originalmente.

6.2 Medidas de avaliação

Tanto na experiência 1 como na avaliação automática realizada na experiência 2 usaram-se as medidas bilingual evaluation understudy (BLEU) e translation error rate (TER), de acordo com o estado da arte [35, 36, 39, 65]. Estas métricas medem a qualidade da tradução com base na comparação do output do sistema, conhecido como **hipótese** e o conjunto de traduções criadas por peritos, chamadas de **referências**. BLEU [66] é calculado com base na correspondência de n-gramas entre a hipótese e as referências. Geralmente, baseia-se na média de unigramas, bigramas, trigramas e tetragramas. Seguindo as experiências conduzidas noutros artigos, calculou-se o BLEU cumulativo até 4-gramas. Os seus valores variam de 0 a 1, sendo 1 uma correspondência exata entre as hipóteses e as referências. Por sua vez, TER [67] é uma extensão de Word Error Rate (WER), medida comummente usada na avaliação de sistemas de reconhecimento automático de fala [68]. TER mede o número de pós-edições (substituições, inserções e eliminações) necessárias para igualar a hipótese à referência, com base na Distância de Levenshtein entre as palavras da hipótese e as da referência. Assim, quanto maior for o valor de TER, mais edições deverão ser feitas à hipótese. O valor total de BLEU corresponde ao valor cumulativo de 4-gramas e o valor total de TER, a média dos valores TER de cada frase. BLEU e TER foram calculados, respetivamente, pela função BLEU_SCORE da biblioteca NLTK[1] e pela biblioteca PYTER[2]. Na avaliação manual da experiência 2 e com base em experiências idênticas noutros artigos [35, 39, 65] recorreu-se a escalas *Mean Opinion Score* (MOS) para classificar a qualidade da tradução.

6.3 Experiência 1: avaliação do módulo de construção das regras

6.3.1 Configuração experimental

As regras de tradução e o dicionário bilingue criados automaticamente no módulo de construção de regras são a base do tradutor. Erros provocados por componentes deste módulo são refletidos na tradução.

Com esta avaliação pretende-se responder às seguintes questões:

1. Quais são os passos deste módulo que originam falhas na construção das regras?
2. Que efeitos negativos exercem esses passos?

As regras automáticas e dicionário bilingue foram criados a partir de 73 frases do corpus anotado. A avaliação deste módulo consiste na tradução destas frases pelo tradutor desenvolvido e na sua

[1] https://www.nltk.org/api/nltk.translate.html
[2] https://github.com/roy-ht/pyter

comparação manual com as respetivas frases em LGP, anotadas no corpus por especialistas. Esta avaliação extrínseca permite avaliar a eficácia do alinhamento. No entanto, as frases em LGP resultantes do tradutor podem ou não ter a estrutura da regra de tradução aplicada na fase da transferência, isto porque passam pela fase de geração, onde são aplicadas regras manuais que alteram a ordem das glosas e acrescentam/eliminam glosas e marcadores conforme as características morfológicas do léxico. Dado que o objetivo é avaliar as componentes do módulo (análise sintática, morfossintática e alinhamento), decidiu-se descartar os processamentos que alteram a estrutura após a aplicação das regras, como as marcações do feminino e de tempos verbais. Assim, os únicos processamentos adicionais consistem nas regras manuais que removem os verbos *Ser* e *Estar*, a conjunção coordenativa copulativa *e* e preposições e datilologia. Por exemplo, ao invés de se analisar a frase em LGP, *PASSADO PORQUE DT(L-U-T-E-R-O) TER REVOLTAR*, na qual a glosa PASSADO e o marcador DT() foram adicionados, analisou-se o output da aplicação de uma regra de tradução, *PORQUE LUTERO TER REVOLTAR*, conservando a estrutura original. As expressões faciais também são removidas por não serem marcadas nas frases em LGP do corpus. Com isto, as frases a avaliar contêm a estrutura resultante da aplicação da regra de tradução que melhor se ajustou. As frases traduzidas são comparadas às respetivas traduções em LGP usando as medidas automáticas TER e BLEU presentes no corpus e as suas diferenças são analisadas manualmente. Das 73 frases do corpus apenas 67 foram traduzidas e avaliadas. As restantes não foram possíveis de traduzir devido a um bug na ferramenta SpaCy que faz com que seja atribuída mais do que uma raiz à árvore sintática da frase (um comportamento não intencional e não documentado).

6.3.2 Resultados

Os valores de TER e BLEU obtidos nesta experiência foram de 0.85 e 0, respetivamente. Estes resultados indicam que existem bastantes diferenças entre as traduções e as respetivas referências. Essas diferenças podem ocorrer a dois níveis, lexical (glosas) e sintático (ordem das glosas). Um dos objetivos desta avaliação é descobrir a origem dessas diferenças e relacioná-las com as componentes do módulo de construção de regras. A comparação das traduções e respetivas referências permitiu reunir um conjunto de possíveis causas dessas divergências.

Para as diferenças lexicais, reuniram-se as seguintes causas:

- A tradução lexical é literal, i.e, o gesto é traduzido para a glosa da palavra quando a palavra não existe no dicionário bilingue. Como as medidas usadas não têm em conta a semelhança semântica entre a tradução e a referência, glosas com o mesmo significado são considerados diferentes. Por exemplo, a tradução e a referência da frase *Foi muito importante.* são, respetivamente, *TER-MUITO IMPORTANTE* e *VALOR TER-MUITO*. Os gestos realçadas são sinónimos, mas as medidas consideram-nos diferentes.

- O alinhamento não é perfeito, nem todos os pares palavra-gesto são considerados uma correspondência. Por exemplo, *sorridentes* não é alinhado com *CARA-SORRIR*. Isto implica que este par não seja uma entrada do dicionário bilingue e, consequentemente, *sorridentes* não é traduzido para *CARA-SORRIR*.

- Algumas frases do corpus são anafóricas, pelo que a frase em LGP possui gestos para se referir a um elemento anteriormente mencionado. A tradução de referência da frase *Porque dentro de Itália o Papa levava uma vida boa, de riqueza.* possui a glosa *ELE* para se referir à entidade *Papa* que foi mencionada na frase anterior.

- E ainda, apesar de pouco frequente, a existência de inconsistências na anotação da tradução para português e da frase em LGP no corpus de referência. Por exemplo, a tradução da frase em LGP *SÉCULO 17 TER-MUITO ARTE IGREJA POLÍTICA SOCIAL DINHEIRO DESENVOLVIMENTO* é anotada como *No século 17 houve um grande desenvolvimento artístico, religioso, político e social.*, na qual não há referência ao conceito *DINHEIRO*.

Quanto à ordem das glosas, esta é definida pelas regras de tradução automáticas, por isso, as divergências sintáticas resultam da aplicação das regras frásicas e morfossintáticas. As diferenças provocadas por estas regras estão ligadas a falhas nas componentes de análise sintática e do alinhamento.

- Verifica-se que na análise sintática perdem-se algumas informações da frase, que podem dever-se a erros na análise de dependências realizada pela ferramenta SpaCy ou a falhas na implementação para identificar os elementos frásicos (sujeito, predicado e modificador de frase). Por exemplo, a tradução da frase *Primeiro em Arte, segundo em coisas que consideravam muito belas.* resulta numa frase incompleta: *PRIMEIRO ARTE DOIS COISAS QUE*. O predicado da frase não foi identificado na análise sintática. Isto implica que não foi estabelecida nenhuma regra morfossintática para o predicado. Este problema não foi detetado durante a implementação, porque as frases testadas eram mais simples (Anexo E).

- O facto do alinhamento não ser totalmente eficaz, não afeta apenas a tradução lexical, mas também a construção de regras automáticas por estar associado a perda de informação.

6.4 Experiência 2: avaliação do módulo de tradução

Após o desenvolvimento do sistema de tradução, impõem-se algumas questões:

1. Quão bem o sistema desenvolvido traduz?

2. Com este procedimento, a semântica da frase em português é preservada?

3. A gramática (léxico, sintaxe e expressões faciais) da LGP é respeitada?

4. Qual é o impacto das regras automáticas no desempenho do sistema?

Para responder a estas questões duas avaliações foram conduzidas, uma automática e outra manual. Na avaliação automática (Secção 6.4.1) são comparados três sistemas, o sistema proposto (Secção 5), sistema baseado apenas nas regras manuais (Secção 6.4.1.C) e o sistema *baseline* (Secção 6.4.1.B). Na avaliação manual (Secção 6.4.2), pediu-se a especialistas em LGP e em linguística para avaliar o *output* do sistema proposto em termos de adequação[3] e fluência[4]. Em ambas as avaliações são usados os dados do corpus de teste descrito na Secção 6.1.2.

6.4.1 Avaliação automática

6.4.1.A Configuração experimental

As 58 frases em português do corpus de teste (Secção 6.1.2) foram traduzidas pelo sistema e o seu resultado foi avaliado usando as medidas BLEU e TER. Com o intuito de perceber o impacto da abordagem seguida na qualidade da tradução, o desempenho do presente sistema é comparado com o do o sistema baseado apenas nas regras manuais (Secção 6.4.1.C) e com o do sistema *baseline* descrito na secção 6.4.1.B.

Apesar de a avaliação automática ser mais barata, consistente e flexível com modificações, as medidas não têm em conta se o significado das frases foi preservado na tradução. Por outro lado, permite que o sistema desenvolvido seja comparado com outros sistemas de traduções.

6.4.1.B Baseline: português gestuado

O sistema *baseline* consiste na produção de português gestuado, ou seja, as frases traduzidas seguem a ordem sintática do português. As frases em português gestuado não apresentam determinantes artigos (definidos e indefinidos), preposições e expressões faciais. Por exemplo, a tradução para português gestuado da frase *Quem comeu o bolo?* é *QUEM COMER BOLO*. O resultado da tradução das 58 frases em português da coleção dourada (Secção 6.1.2) por este sistema baseline encontra-se no Anexo F.

6.4.1.C Sistema baseado apenas em regras manuais

A diferença entre este sistema de tradução e o PE2LGP está na transferência da estrutura sintática. Neste sistema, esta é realizada somente através da aplicação das regras frásicas e das regras manuais,

[3] Em inglês *adequacy*, permite saber se o significado da frase na língua origem é totalmente expresso na frase na língua alvo.
[4] Em inglês *fluency*, mede a veracidade da gramática juntamente com a facilidade de compreensão da tradução.

enquanto que no PE2LGP realiza-se com a aplicação de todas as regras: regras morfossintáticas, frásicas e manuais.

6.4.1.D Configurações

É comum na avaliação automática comparar-se o sistema desenvolvido com outros sistemas. Por isso, as traduções realizadas por este sistema, pelo sistema baseline (na Secção 6.4.1.B) e pelo sistema baseado apenas nas regras manuais (Secção 6.4.1.C) serão comparadas para avaliar a qualidade do sistema desenvolvido.

As frases em LGP que saem do sistema desenvolvido e do sistema baseado apenas nas regras manuais podem seguir duas estruturas frásicas distintas (tendo em conta os dados usados), a ordem SOV, que é a tradicional, e a ordem mais frequente do corpus. Por isso, avaliaram-se estas duas estruturas de outputs em ambos os sistemas.

Além disso, para ambas as configurações anteriores, decidiu-se comparar os resultados da tradução das frases tendo e não tendo em conta as expressões faciais. Por exemplo a frase em 31 *AMANHÃ DT(C-A-R-O-L-I-N-A) VESTIR NÃO(headhake)*, sem expressões faciais ficaria *AMANHÃ DT(C-A-R-O-L-I-N-A) VESTIR NÃO*.

Assim, no total conduziram-se 10 experiências, dispostas na Tabela 6.1.

As configurações I e II são do sistema baseline, as configurações III, IV, V e VI pertencem ao sistema baseado apenas nas regras manuais e formam o **conjunto 1**, por fim, as configurações VII, VIII, IX e X são do sistema proposto (PE2LGP) e formam o **conjunto 2**.

Configuração	**Procedimento**
I	Baseline sem expressão facial
II	Baseline com expressão facial
	Conjunto 1
III	Estrutura SOV sem expressão facial
IV	Estrutura SOV com expressão facial
V	Estrutura de acordo com as regras frásicas e sem expressão facial
VI	Estrutura de acordo com as regras frásicas com expressão facial
	Conjunto 2
VII	Estrutura SOV sem expressão facial
VIII	Estrutura SOV com expressão facial
IX	Estrutura de acordo com as regras e sem expressão facial
X	Estrutura de acordo com as regras e com expressão facial

Tabela 6.1: Configurações experimentais.

6.4.1.E Resultados

A Tabela 6.2 apresenta os resultados para as medidas TER e BLEU das configurações dos vários sistemas. No Anexo G encontram-se os valores de TER para cada frase na configuração X, que é a configuração cuja tradução é baseada em todas as regras (regras automáticas e manuais).

Configuração	TER	BLEU			
		1-grama	2-gramas	3-gramas	4-gramas
I	0.69	0.66	0.17	0	0
II	0.86	0.5	0.13	0	0
Conjunto 1					
III	0.28	0.79	0.66	0.58	0.42
IV	0.3	0.75	0.64	0.55	0.38
V	0.39	0.79	0.5	0.37	0.23
VI	0.4	0.75	0.47	0.35	0.21
Conjunto 2					
VII	0.29	0.78	0.66	0.56	0.39
VIII	0.29	0.77	0.64	0.54	0.37
IX	0.41	0.76	0.51	0.38	0.24
X	0.4	0.77	0.49	0.36	0.21

Tabela 6.2: Resultados das 10 configurações experimentais.

Os resultados entre as diferentes configurações dos conjunto 1 e 2 são semelhantes mas os melhores resultados foram obtidos para as traduções baseadas apenas nas regras manuais, seguindo a estrutura SOV e sem expressões faciais (configuração III), alcançando 0.28 para TER e 0.42 para BLEU. O sistema baseline, pelo contrário, apresenta para as duas configurações (com e sem expressão facial) os piores resultados.

A – Sistema baseline Vs PE2LGP 4.0

Os resultados do sistema desenvolvido superaram os do sistema baseline, atingindo 0.29 de TER e 0.37 de BLEU para a estrutura SOV com expressões faciais (configuração VIII). Estes valores mostram que a aplicação das regras automáticas e das regras manuais na transferência gramatical melhoram consideravelmente a qualidade das traduções, produzindo LGP e não somente português gestuado.

B – Conjunto 1 Vs Conjunto 2

Os resultados entre as configurações que pertencem ao conjunto 1 e àquelas que pertencem ao conjunto 2 apresentam ligeiras diferenças. Por exemplo, os valores de TER e BLEU entre a configuração III e a configuração VII têm uma diferença de 0.01 e de 0.03, respetivamente. Apenas 8 das 58 traduções com o sistema PE2LGP apresentam diferenças em relação às traduções com o sistema

baseado somente em regras manuais. Estas diferenças serão analisadas nesta secção. Na Tabela 6.3 listam-se as 8 sequências de glosas resultantes dessas duas traduções.

Par de frases	Sistema baseado apenas em regras manuais	Sistema PE2LGP
1	NAMORADO MEU TER OLHOS VERDES	NAMORADO MEU TER VERDES OLHOS
2	ALI POUCO SOL	ALI ALI POUCO SOL
3	MULHER MENINO ÓCULOS VER FLOR {NÃO}(headshake)	ÓCULOS MULHER MENINO VER FLOR {NÃO}(headshake)
4	ALI POUCO SOL {NÃO}(headshake)	ALI POUCO SOL {NÃO}(headshake) {NÃO}(headshake)
5	{SEGURANÇA QUER RESPEITO}(q)	{SEGURANÇA RESPEITO QUER}(q)
6	{PAPA BOM OUVINTE}(q)	{PAPA OUVINTE BOM}(q)
7	{ELA CASA AZUL TER}(q)	{ELA AZUL CASA TER}(q)
8	{ALI POUCO SOL}(q)	{ALI SOL POUCO}(q)

Tabela 6.3: As 8 sequências de glosas que são diferentes entre os dois sistemas.

Ao comparar as traduções dos dois sistemas na Tabela 6.3, verifica-se que as regras morfossintáticas alteram a ordem dos adjetivos e a marcação dos advérbios. Nos pares de frases 1, 6 e 7, os adjetivos (VERDES, BOM e AZUL) são colocados antes dos substantivos. Nos pares de frases 2 e 4, há uma repetição dos advérbios *ALI* e *NÃO*. Se a frase em 4 é a negação da frase em 2, então porque é que o advérbio repetido não é o mesmo? Para saber-se as origens desta incoerência, analisaram-se os resultados de cada etapa da tradução dessas frases. Para ambas, *Ali* e *Ali não* são considerados modificadores e *está pouco sol*, o predicado. As regras escolhidas também foram as mesmas, ADV1 $\rightarrow$ ADV1 ADV1 para os modificadores e V1 N2 $\rightarrow$ V1 N2 para os predicados. A razão da incoerência foi encontrada nas operações de edição dadas pela Distância de Edição, contudo a origem não é essa. O problema surge porque as etiquetas das regras não discriminam a subclasse dos constituintes, o que leva a que não haja distinção entre o advérbio de negação (*não*) e o advérbio *ali*, são ambos vistos como advérbios pelo algoritmo. Para que o lado português da regra *ADV1* fique igual à estrutura sintática do modificador *Ali não*, seria necessário inserir um advérbio de negação depois do ADV1, mas como não há distinção entre advérbios de negação de outros tipos de advérbios, as operações resultantes do algoritmo indicam uma inserção de um advérbio antes do ADV1. Isto implica que a regra aplicada corresponda a ADV2 ADV1 $\rightarrow$ ADV2 ADV1 ADV1. Enquanto que na primeira frase, ADV1 corresponde a *Ali*, na segunda corresponde ao *não*.

As trocas da ordem das glosas nas restantes frases devem-se à aplicação das regras morfossintáticas. Para a versão interrogativa da frase anterior, *Ali está pouco sol?*, as regras escolhidas são diferentes devido ao tipo da frase ser diferente. Para o modificador a regra aplicada foi ADV1 $\rightarrow$ ADV1 e para o predicado a regra aplicada foi V1 DET3 N2 $\rightarrow$ N2 V1 DET3, o que justifica a troca da posição do gesto *SOL* para antes do gesto *POUCO*.

Relativamente ao par de frases 5, são traduções da frase *O segurança quer respeito?*. A ferramenta classificou incorretamente um constituinte, o verbo *quer* foi classificado como uma conjunção coordenativa. Além de influenciar a escolha da regra, a palavra *quer*, ao invés de ser representado por *QUERER*, é representada por *QUER*.

Com a aplicação das regras automáticas foram aperfeiçoadas as traduções de duas frases inter-

rogativas, igualando-as à referência. São elas *O papa é bom ouvinte?* e *Ali está pouco sol?*. Na última frase, houve um erro na classificação morfossintática da palavra *pouco*, contudo esse erro não influenciou a escolha da regra.

A proximidade entre os valores dos dois conjuntos deve-se a duas razões. A maioria das regras morfossintáticas aplicadas às frases declarativas afirmativas e negativas não alteram a estrutura sintática da frase, i.e, a ordem dos constituintes do lado português da regra é igual à ordem do lado da LGP. Quanto à tradução de interrogativas, as regras frásicas aplicadas definem que o verbo deverá ser marcado no final da frase em LGP. Dado que estas são aplicadas em ambos os conjuntos, então, a maioria das traduções de ambos os conjuntos é igual. Além disso, dado que a estrutura sintática da maioria das frases do corpus do teste são semelhantes, as regras aplicadas são as mesmas, principalmente para o sujeito. Com estes resultados não é possível tirar conclusões sobre o impacto das regras automáticas no desempenho do sistema de tradução. Uma avaliação futura com um corpus de teste com maior variabilidade de estruturas poderá responder a essa pergunta.

Em ambos os conjuntos, as traduções com as estruturas SOV (configurações III, IV, VII e VIII) são as que apresentam resultados mais altos, porque as referências no corpus de teste seguem a ordem frásica SOV.

No anexo G estão listados os valores de TER para cada frase para a configuração X. 5 das 58 traduções obtiveram um valor 1 de TER, o que indica algum trabalho de edição para que as traduções fiquem iguais à referência. Uma das traduções é particularmente interessante por ilustrar não só o contraste entre a LP e a LGP como também as limitações da avaliação automática. A tradução pelo sistema da frase *Ele é espanhol.* é *ELE ESPANHOL* mas a sua referência é *ESPANHA PAÍS DELE*. No entanto, a avaliação manual desta frase indica que a tradução do sistema está também correta.

C – Configurações com expressão facial Vs configurações sem expressão facial

Não existe uma diferença significativa entre os valores das traduções com expressão facial (IV, VI, VIII e X) e das sem expressão facial (III, V, VII e IX). A maioria das traduções com expressão facial apresentam resultados mais baixos do que aquelas que não têm expressão facial.

Das 58 frases, 20 têm expressões faciais ligadas à negação e as outras 20 à interrogação. A marcação das expressões faciais nas frases interrogativas vai ao encontro da marcação realizada nas referências, contudo a marcação das expressões nas frases negativas implementada no tradutor não corresponde às das referências. Apenas 5 das 20 referências seguem a regra implementada. Na maioria das referências a negação do verbo é realizada através da expressão facial encher as bochechas e abanar a cabeça simultaneamente ao verbo, como na referência *MULHER MENINO ÓCULOS FLOR* {*VER*}*(NÃO)*, noutras a marcação é feita pela adição dessa expressão facial depois do verbo, como

em *MULHER CEGA SIMPÁTICA (NÃO)* e ainda pela adição de gestos manuais como *THU* e *NÃO* acompanhados pela expressão facial *headshake* no final da frase, como em *NETO FORMIGA COMER THU(headshake)* e *CRIANÇA INTELIGENTE NÃO(headshake)*, entre outras.

6.4.2 Avaliação manual

6.4.2.A Configuração experimental

As métricas BLEU e TER usadas na avaliação automática não refletem a qualidade semântica da tradução produzida pelo sistema [68, 69]. Para a determinar é necessário o conhecimento humano, sendo importante realizar uma avaliação com utilizadores, cujos resultados são mais informativos [35]. O objetivo desta avaliação é saber se o significado da frase em português prevalece na hipótese, mesmo havendo diferenças na gramática e léxico em relação à referência. Assim escolheram-se frases da avaliação automática que apresentaram diferenças em relação à referência.

A avaliação foi realizada com 4 peritos em linguística e com conhecimentos de LGP, através de uma entrevista individual por vídeochamada, na qual se pedia para classificarem a qualidade da tradução de 11 frases, quer a nível gramatical quer a nível semântico. Antes da aplicação do questionário final, realizou-se um teste piloto para ajustar a entrevista.

6.4.2.B Entrevista

Dado o número reduzido de participantes, optou-se por realizar uma entrevista, de forma a captar dados relevantes através do discurso e do comportamento dos participantes, que não seria possível através de questionários, cujos dados seriam quantitativamente pouco significantes. As entrevistas foram individuais, mas iguais para todos os participantes, e concretizaram-se através de vídeo-chamadas por *Zoom*. A duração média das entrevistas foi de 40 minutos. De modo a que o entrevistador tenha mais controlo sobre a entrevista, o questionário foi partilhado por ecrã e preenchido pelo entrevistador. Conforme as respostas do participante, pediu-se para justificá-las e foram sendo feitas outras perguntas conforme o decorrer da entrevista.

A cada frase estão associadas 3 perguntas. Na primeira apresenta-se uma sequência de glosas e é pedido ao participante para dizer a respetiva tradução em português. As duas restantes questões baseiam-se na análise da mesma sequência de glosas e da tradução em português de referência. Na segunda questão pede-se ao participante para classificar a qualidade da tradução em *pobre*, *justo* e *bom*, em que pobre corresponde a uma tradução incorreta (o significado da tradução está incorreto), justo para os casos em que o significado da tradução é o correto mas a gramática falha em alguns aspetos e bom quando o significado da tradução e a gramática estão corretos. Esta escala de três valores é comum em avaliações manuais de sistemas de tradução [35, 65]. Por último, é pedido aos participantes

para avaliar aspetos linguísticos e gramaticais da tradução, nomeadamente a expressão facial, ordem dos constituintes sintáticos e o léxico, com base na escala *Incorreto*, *Parcialmente (in)correto* e *Não se aplica*. Esta questão permite justificar o voto realizado na pergunta anterior sobre a qualidade da tradução. Um extracto do questionário (para uma frase) encontra-se no Anexo H.

6.4.2.C Dados de teste

Tendo em conta o objectivo mencionado anteriormente, as 11 sequências de glosas usadas na avaliação manual foram escolhidas começando por excluir aquelas com valor TER (conhecido da avaliação automática da experiência com a estrutura de acordo com as regras de tradução e com expressão facial) igual a 0, por serem exactamente iguais à referência. BLEU é uma métrica de corpus e não de frases individuais [68], por isso os seus valores não fazem parte do critério de escolha das frases em LGP. De entre as restantes, foram escolhidas para a avaliação manual aquelas com maior potencial para produzir resultados interessantes, ou seja aquelas que possuem diferenças significativas de léxico e ordem das glosas que poderão afetar a compreensão da frase. Por exemplo, a frase *O estado tem o poder?* é traduzida como *ESTADO PODER TER(q)*, enquanto que a referência é *ESTADO PODER HÁ(q)*, pelo que existe uma diferença entre a referência e a tradução no verbo, mas será que afeta a compreensão da frase? No Anexo I listam-se as frases escolhidas e as respetivas traduções do presente sistema.

6.4.2.D Teste piloto

Antes de se aplicar o teste final, um teste preliminar foi conduzido com o propósito de avaliar a exequibilidade e adequação do questionário/entrevista em relação à sua duração, à sua estrutura e questões, tendo em conta o objetivo da avaliação. O questionário foi ajustado, de acordo com as indicações resultantes da simulação da entrevista. O teste piloto foi realizado com a intérprete que criou a coleção dourada usada na avaliação (Secção 6.1.2), não integrando a amostra final. Ao fim de 40 minutos, o teste terminou, contudo, é importante realçar que, pelo facto da intérprete ter criado as frases em estudo, a interpretação das sequências de glosas foi mais rápida. Assim, uma das alterações feitas foi reduzir o tempo do questionário, ou seja, diminuiu-se o número de frases a analisar. Inicialmente, o questionário era composto por 16 frases, reduziu-se para 11. Durante o teste, foram encontrados mais dois aspetos a melhorar:

- Não era claro no questionário que as expressões faciais marcadas nas sequências de glosas são apenas referentes a expressões gramaticais (que marcam frases negativas e interrogativas) e não às lexicais, pelo que esta informação foi explicitada na introdução do questionário.
- Dado que o questionário é mostrado ao avaliador mas preenchido pelo entrevistador, os avaliadores não têm a liberdade de navegar no questionário como teriam se estivesse nas suas mãos.

Isto fez com que, numa questão em particular, a avaliadora tivesse que pedir várias vezes para o entrevistador voltar para trás para poder rever uma informação. Assim, repetiu-se, antes da pergunta, a informação relevante para a sua resposta.

6.4.2.E Participantes

As questões feitas aos participantes ou avaliadores sobre as sequências de glosas são de carácter linguístico, pelo que os mesmo deverão ter conhecimentos de linguística, além de estarem familiarizados com glosas e terem conhecimentos de LGP. Cumprindo este critério, conseguiu-se 4 avaliadores com diferentes níveis de LGP: 1 nativo (surdo), 1 fluentes e 2 com nível intermédio na língua.

6.4.2.F Resultados

A qualidade da tradução do presente sistema para 25% das frases foi *justa*, enquanto que para as restantes (75%) foi classificada como *boa*. Estes valores indicam que o significado da frase é preservado em todas as traduções do sistema PE2LGP e em 75% das traduções, a gramática está também correta. Para 4 das traduções com os piores valores de TER na avaliação automática foi atribuído unanimemente pelos participantes o valor *Bom* para a qualidade das traduções.

Nesta secção são analisados os erros gramaticais existentes nas traduções.

Ordem frásica e ordem das glosas

Estes são os resultados sobre as ordens frásicas e dos constituintes morfossintáticos:

- 82% das ordens frásicas das traduções estão corretas, enquanto que 18% apresentam erros na ordem.
- 77% das traduções respeitam a ordem dos constituintes morfossintáticos e 23% possui erros nessa ordem.

As frases que apresentam uma ordem frásica errada, apresentam erros na ordem das glosas. Os resultados indicam que as frases negativas são as que possuem mais erros na ordem frásica e morfossintática. Para todos os participantes o verbo *TER* na frase *NAMORADO MEU TER OLHOS VERDES* $\{$*NÃO*$\}$*(headshake)* deveria ser colocado antes do gesto NÃO, pois a negação é sobre o verbo. A ordem frásica seria, então, SOV e não SVO. Este padrão foi encontrado em todas as frases negativas. Ainda sobre esta frase, o verbo *TER* foi considerado por 50% dos participantes, curiosamente aqueles com melhor níveis na LGP como um verbo copulativo, ou seja, deverá estar incorporado no objeto, não sendo representado em LGP, ficando assim *NAMORADO MEU OLHOS VERDES* $\{$*NÃO*$\}$*(headshake)*.

1 dos participantes referiu que o adjetivo *VERDES* deveria estar antes do nome *OLHOS*. Semelhante à frase anterior, o verbo da frase *ELES GOSTAR MASSA {NÃO} (headshake)* deveria ser colocado antes do gesto *NÃO*, de acordo com os participantes.

Léxico

Os resultados indicam que 82% das traduções possuem as glosas corretas.

Uma das limitações do uso das glosas é a ambiguidade lexical que elas produzem, uma mesma glosa pode estar associada a vários gestos. Este problema dificultou a compreensão da sequência de glosas *SEGURANÇA QUERER TAMBÉM RESPEITO*, na qual a maioria dos participantes interpretou *SEGURANÇA* como o sentimento de segurança e não a profissão de segurança. Para distinguir estes dois casos, sugeriram a adição dos gestos *HOMEM* ou *PESSOA* antes da glosa *SEGURANÇA*. O facto de serem frases soltas, sem contexto, dificulta a desambiguação estes casos.

Outros erros do léxico relacionam-se com a marcação das frases negativas. Os gestos manuais usados na marcação da negação do verbo em LGP são variados e dependem do verbo, no entanto não existe um consenso sobre a marcação da negação de uma mesma frase entre os participantes. A marcação da negação na frase *ELES GOSTAR MASSA {NÃO} (headshake)* foi considerada errada por 50% dos participantes. Estes defenderam que a negação, nesta frase especifica é simultânea ao verbo *GOSTAR* e faz-se somente por expressão facial e não pelo gesto manual *NÃO*. Noutras frases, o gesto manual *NÃO* foi considerado inadequado para marcar a negação naquele contexto, deveria ser o gesto manual *NÃO-HÁ*.

Por fim, para a última frase, *EU PERGUNTAR TU NÃO(headshake)* foram apontados vários problemas. A maioria dos participantes (75%) disse que a frase em português desta sequência de glosas era *Eu não te perguntei.* porque numa conversa é incomum pronunciar-se a frase *Eu não te pergunto.*. Mais uma vez, o facto de as frases não terem contexto, dificultou a tradução para português das sequências de glosas.

Expressão facial e a sua duração

Os resultados obtidos sobre a marcação das expressões faciais indicam que:

- 79% das traduções tem a expressão facial correta (21% incorretas).
- 71% da duração das expressões faciais está correta (29% não acompanham os gestos certos).

Nas duas frases interrogativas, a marcação das expressões faciais foi classificada como correta, contudo, os participantes indicaram que existem outras possibilidades que para eles são as mais corretas. Essas possibilidades variam entre os participantes, não havendo um consenso entre as opiniões.

Por exemplo, para a frase *ESTADO PODER TER ?*, foram indicadas as seguintes variações da posição da expressão facial interrogativa (levantar o queixo, inclinar a cabeça para trás e franzir as sobrancelhas) na frase: ocorre na última glosa (TER) ou a partir da glosa *PODER* até ao final da frase.

Mais uma vez as falhas das expressões faciais estão na marcação da negação, não havendo concordância entre os participantes. As observações de alguns participantes indicam que a expressão *headshake* nem sempre é usada na marcação da negação. Existem várias expressões faciais para marcar a negativa que depende dos verbos e outras características gramaticais presentes na frase.

6.5 Discussão

Os resultados da experiência 1 indicam que o módulo de construção das regras apresenta algumas limitações, nomeadamente na identificação dos constituintes frásicos (análise sintática) e no alinhamento. Estas limitações afetam a caracterização dos fenómenos gramaticais presentes numa frase, nas regras automáticas.

Os resultados da experiência 2 mostram que a aplicação das regras automáticas e de regras manuais na transferência gramatical melhoram consideravelmente a qualidade das traduções face ao sistema baseline. As traduções de frases negativas foram as que obtiveram piores resultados, a marcação da negativa depende de aspetos gramaticais na frase, nomeadamente do verbo. Os resultados da avaliação manual, para uma mostra pequena de frases, indicam que apesar de existirem alguns erros nas traduções, estes não dificultam a compreensão da mensagem.

Por fim, a comparação dos resultados da experiência 1 com os da experiência 2 indica que a análise sintática apresenta mais falhas em frases maiores cujas palavras possuem relações de dependências mais complexas.

7

Conclusão

Conteúdo

7.1 Conclusões

A construção de um sistema de tradução de LP para LGP baseado em regras manuais é condicionada pelos poucos recursos disponíveis sobre a gramática da LGP e para o processamento de texto em português.

A Universidade Católica Portuguesa está a desenvolver o primeiro corpus de português europeu e de língua gestual portuguesa anotado com informações gramaticais das frases em LGP, que abrangem vários fenómenos linguísticos. A principal inovação deste tradutor face aos seus antecessores é a exploração do novo corpus. Por norma, os tradutores desenvolvidos anteriormente utilizam exclusivamente regras de tradução manuais. Assim, o sistema de tradução apresentado, além de regras manuais, faz uso deste corpus anotado para gerar regras de tradução automáticas com o objetivo de obter traduções de português para LGP que reflitam a gramática da língua.

Os resultados mostram que a abordagem de tradução seguida é capaz de captar fenómenos gramaticais e produzir frases em LGP ao invés de português gestuado. O sistema mostrou bons resultados a nível da inteligibilidade, apesar das conhecidas limitações na marcação da negação, identificação dos elementos frásicos, na análise morfossintática e na transferência sintática.

7.2 Trabalho futuro

O sistema proposto é uma primeira versão de um tradutor de português para LGP baseado em informações gramaticais extraídas de um corpus. Este sistema integra-se num projeto maior, em que um dos principais objetivos é a criação de um tradutor automático de português para LGP, no qual as frases são produzidas por um avatar.

Apesar de os resultados obtidos serem bons, existem alguns aspetos a melhorar e a estender. Alguns desses aspetos relacionam-se com a ligação entre a sequência de glosas e a produção dos gestos pelo avatar. Estes são alguns pontos identificados:

1. **Formalismo gramatical**: a disposição das regras de tradução numa gramática formal síncrona, como as *Synchronous tree-adjoining grammars* (Secção 3.1.1) permitirá caracterizar melhor os diferentes fenómenos gramaticais da língua.

2. **Regras morfossintáticas mais finas**: discriminar nas regras morfossintáticas também as subclasses gramaticais permitiria ter regras com maior detalhe, ajudando a evitar que as regras sejam usadas em casos onde não deveriam ser aplicadas.

3. **Marcação da negação**: identificar e implementar os gestos manuais e não manuais mais adequados para marcar a negação de uma dada frase.

4. **Alinhamento**: melhorar o algoritmo do alinhamento.

5. **Análise sintática**: melhorar a identificação dos elementos frásicos (sujeito, predicado e modificador de frase).

6. **Processamento de particularidades morfossintáticas da LGP**:

   - **Classificadores**: o corpus de referência poderá ser útil na identificação dos classificadores.
   - **Negação incorporada**: o sistema implementado não reconhece verbos com negação incorporada, como *NÃO-QUERER*, que é diferente do gesto *QUERER*.
   - **Preposições**: as preposições são incorporadas no movimento dos gestos para identificar, por exemplo, os locais inicial e final do objeto que se move. É preciso decidir uma notação para identificar as diferentes preposições e os movimentos que advêm delas.
   - **Verbos de concordância**: nas línguas gestuais existem verbos cuja trajetória, movimento e/ou orientação são alterados consoante a posição dos argumentos interno e externos, e esses argumentos são omitidos lexicalmente e incorporados no movimento de trajetória do verbo [70]. Por exemplo, a produção em LGP da frase *Eu dou-te.* não é *EU DAR TU*, como o tradutor traduziria mas é apenas um gesto com posição inicial no *EU* (na pessoa que está a gestuar) e posição final no *TU*. Portanto, além da sua identificação é necessário criar convenções para representá-los na sequência de glosas.

7. **Ferramentas em falta**:

   - Dicionário bilingue de português europeu e LGP com correspondências entre gestos em glosa e palavras.
   - Ferramenta de conversão de uma palavra com certa classe gramatical para a palavra correspondente de outra classe gramatical. Por exemplo, passar do adjetivo *inteligente* para o nome *inteligência*.

8. **Ambiguidade lexical**: as glosas podem estar associadas a diferentes gestos dependendo do contexto onde o gesto se insere. Por exemplo, a glosa *GRANDE* está associada a gestos diferentes dependendo do objeto, se o tamanho do objeto variar na vertical (em altura) então, normalmente, recorre-se ao gesto da alínea a) da Figura 7.1, se o tamanho variar na horizontal (em largura) usa-se o gesto na alínea b).

(a) Gesto *GRANDE* em altura. **(b)** Gesto *GRANDE* em largura.

Figura 7.1: Exemplo de ambiguidade lexical da glosa *GRANDE*. Estas imagens foram retiradas do dicionário de línguas gestuais *spread the sign*.

9. **Construção de mais regras**: quanto mais regras melhor, pois evita que fenómenos gramaticais excecionais não sejam considerados a regra apropriada.

10. **Aprendizagem automática**: treinar um modelo probabilístico com o corpus, quando este for suficientemente grande.

11. **Prosódia**: a sequência de glosas do tradutor não identifica elementos de entoação, pausas de discursos, etc. Criar convenções de pontuação para explicitar a prosódia e glosa.

Hoje em dia, a área de NLP encontra-se em expansão, principalmente em línguas populares como o inglês e o chinês. Línguas em minoria, com poucos falantes ou ameaçadas acabam por ficar para trás neste desenvolvimento, quando muitas dessas línguas são de comunidades mais isoladas e que precisam de meios para comunicar. As línguas gestuais são um exemplo desta realidade, porque além de constituírem comunidades com poucos falantes, têm a desvantagem acrescida de os seus falantes terem limitações físicas na utilização de línguas orais. Contudo, tem crescido cada vez mais o interesse no estudo das línguas gestuais, impulsionando a criação de recursos linguísticos e de ferramentas, como tradutores automáticos mesmo face a condições que não são as ideais. No caso da LGP, uma das inovações desta área em crescimento é o corpus anotado da Universidade Católica Portuguesa, um recurso imprescindível não só para a compreensão da língua e estabelecimento de uma gramática, como para a criação de ferramentas computacionais para o seu processamento automático. O sistema de tradução proposto nesta dissertação, em conjunto com todas as contribuições secundárias, constitui o primeiro passo para alavancar o potencial deste corpus na área da tradução automática, estabelecendo a fundação para os futuros desenvolvimentos que permitirão à língua gestual portuguesa ter à sua disposição recursos equiparáveis aos das principais línguas do mundo.

Bibliografia

[1] M. F. da Silva Bettencourt, "A ordem de palavras na língua gestual portuguesa: Breve estudo comparativo com o português e outras línguas gestuais," Master's thesis, Faculdade de Letras da Universidade do Porto, 2015.

[2] I. Almeida, "Exploring Challenges in Avatar-based Translation from European Portuguese to Portuguese Sign Language," Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, 2014.

[3] R. dos Santos, "Pe2lgp: do texto à língua gestual," Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, 2016.

[4] L. Gaspar, "IF2LGP-Intérprete automático de fala em língua portuguesa para língua gestual portuguesa," Master's thesis, Instituto politécnico de Leiria, Leiria, 2015.

[5] R. Ferreira, "Pe2lgp 3.0: from european portuguese to portuguese sign language," Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, 2016.

[6] P. Escudeiro, N. Escudeiro, R. Reis, J. Lopes, M. Norberto, A. B. Baltasar, M. Barbosa, and J. Bidarra, "Virtual sign–a real time bidirectional translator of portuguese sign language," *Procedia Computer Science*, vol. 67, pp. 252–262, 2015.

[7] A. Mineiro and D. Colaço, *Introdução à Fonética e Fonologia na LGP e na língua Portuguesa*. Universidade Católica Editora, 2010.

[8] J. Stokoe, William C., "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf," *The Journal of Deaf Studies and Deaf Education*, vol. 10, pp. 3–37, 2005. [Online]. Available: https://dx.doi.org/10.1093/deafed/eni001

[9] A. da Rocha Costa and G. Dimuro, "Signwriting and swml: Paving the way to sign language processing," *Atelier Traitement Automatique des Langues des Signes, TALN 2003*.

[10] T. Hanke, "Hamnosys-representing sign language data in language resources and language processing contexts," in *LREC*, vol. 4, 2004, pp. 1–6.

[11] R. A. F. Rodrigues, “Compreensão da língua gestual portuguesa em crianças surdas. proposta de um instrumento de avaliação,” Ph.D. dissertation, 2018.

[12] S. Nascimento and M. Correia, *Um olhar sobre a morfologia dos gestos*. Universidade Católica Editora, 2011, vol. 15.

[13] A. P. d. A. Sousa, “Interpretação da língua gestual portuguesa,” Ph.D. dissertation, 2012.

[14] M. Martins and A. I. Mata, “Conexões interfrásicas manuais e não-manuais em lgp: Um estudo preliminar,” *Linguística: Revista de Estudos Linguísticos da Universidade do Porto*, vol. 11, pp. 119–138, 2017.

[15] H. Carmo, V. M. da Silva, and E. Martins, “Os verbos em negação na língua gestual portuguesa,” *Cadernos de Saúde*, vol. 9, pp. 15–25, 2017.

[16] H. d. Carmo *et al.*, “Uma primeira abordagem aos classificadores da língua gestual portuguesa,” Master’s thesis, Universidade Católica Portuguesa, Lisboa, 2016. [Online]. Available: http://hdl.handle.net/10400.14/22600

[17] P. M. Lewis and R. E. Stearns, “Syntax directed transduction,” in *7th Annual Symposium on Switching and Automata Theory (swat 1966)*, Oct 1966, pp. 21–35.

[18] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2018.

[19] D. Gildea and G. Satta, “Synchronous Context-Free Grammars and Optimal Parsing Strategies,” *Computational Linguistics*, vol. 42, no. 2, pp. 207–243, 2016. [Online]. Available: https://doi.org/10.1162/COLI{_}a{_}00246

[20] D. Chiang and K. Knight, “An introduction to synchronous grammars,” *Tutorial available at http://www. isi. edu/ chiang/papers/synchtut. pdf*, 2006.

[21] V. K. Menon and and, “A synchronised tree adjoining grammar for english to tamil machine translation,” in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Aug 2015, pp. 1497–1501.

[22] A. Joshi and O. Rambow, “A formalism for dependency grammar based on tree adjoining grammar,” in *Proceedings of the Conference on Meaning-text Theory*, 2003, pp. 207–216.

[23] S. M. Shieber and Y. Schabes, “Synchronous tree-adjoining grammars,” in *Proceedings of the 13th conference on Computational linguistics-Volume 3*. Association for Computational Linguistics, 1990, pp. 253–258.

[24] A. Joshi and Y. Schabes, “Tree-adjoining grammars,” *Handbook of formal languages*, pp. 69–123, 1997.

[25] J. Earley, “An efficient context-free parsing algorithm,” *Commun. ACM*, vol. 13, pp. 94–102, 1970. [Online]. Available: http://doi.acm.org/10.1145/362007.362035

[26] Y. Schabes and A. K. Joshi, “An earley-type parsing algorithm for tree adjoining grammars,” in *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics*, ser. ACL ’88. Stroudsburg, PA, USA: Association for Computational Linguistics, 1988, pp. 258–269. [Online]. Available: https://doi.org/10.3115/982023.982055

[27] P. Paroubek, Y. Schabes, and A. Joshi, “Xtag - a graphical workbench for developing tree-adjoining grammars,” in *ANLP*, 1992.

[28] K. Arora and S. Agrawal, “Comparative analysis of phrase based, hierarchical and syntax based statistical machine translation.” NISCAIR-CSIR, India, 2018.

[29] M. Costa-jussa, M. Farrús, J. B. Marino, and J. Fonollosa, “Study and comparison of rule-based and statistical catalan-spanish machine translation systems,” *Computing and Informatics*, vol. 31, pp. 245–270, 01 2012.

[30] M. Brour and A. Benabbou, “Atlaslang mts 1: Arabic text language into arabic sign language machine translation system,” *Procedia computer science*, vol. 148, pp. 236–245, 2019.

[31] R. San-Segundo, R. Barra-Chicote *et al.*, “A Spanish speech to sign language translation system for assisting deaf-mute people.” in *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP*, vol. 3, 2006.

[32] M. Davydov and O. Lozynska, “Information system for translation into ukrainian sign language on mobile devices,” in *2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, vol. 1. IEEE, 2017, pp. 48–51.

[33] T. Araújo, F. L. S. Ferreira, D. A. N. S. Silva, L. D. Oliveira, E. D. L. Falcão, L. Domingues, V. Martins, I. A. C. Portela, Y. S. Nóbrega, H. R. G. Lima, G. L. de Souza Filho, T. Tavares, and A. Duarte, “An approach to generate and embed sign language video tracks into multimedia contents,” *Inf. Sci.*, vol. 281, pp. 762–780, 2014.

[34] L. Zhao, K. Kipper *et al.*, “A Machine Translation System from English to American Sign Language,” 2000, pp. 191–193.

[35] H. Luqman and S. A. Mahmoud, “Automatic translation of Arabic text-to-Arabic sign language,” *Universal Access in the Information Society*, jun 2018. [Online]. Available: https://doi.org/10.1007/s10209-018-0622-8

[36] J. Porta, F. López-Colino *et al.*, "A Rule-based Translation from Written Spanish to Spanish Sign Language Glosses," *Comput. Speech Lang.*, vol. 28, no. 3, pp. 788–811, may 2014. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2013.10.003

[37] A. Othman and M. Jemni, "Statistical Sign Language Machine Translation: from English written textto American Sign Language Gloss," *International Journal of Computer Science Issues*, vol. 8, pp. 65–73, 2011.

[38] J. Bungeroth and H. Ney, "Statistical sign language translation," *Workshop on Representation and Processing of Sign Languages, 4th International Conference on Language Resources and Evaluation, LREC 2004*, vol. 4, pp. 105–108.

[39] H. Su and C. Wu, "Improving Structural Statistical Machine Translation for Sign Language With Small Corpus Using Thematic Role Templates as Translation Memory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1305–1315, 2009.

[40] Y.-H. Chiu, C.-H. Wu, H.-Y. Su, and C.-J. Cheng, "Joint optimization of word alignment and epenthesis generation for chinese to taiwanese sign synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 28–39, 2007.

[41] I. Almeida, L. Coheur, and S. Candeias, "From european portuguese to portuguese sign language," in *Proceedings of SLPAT: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 01 2015, pp. 140–143.

[42] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.

[43] J. Ferreira, H. Gonçalo Oliveira, and R. Rodrigues, "Improving nltk for processing portuguese," in *8th Symposium on Languages, Applications and Technologies (SLATE 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[44] R. Rodrigues, H. Gonçalo Oliveira, and P. Gomes, "Nlpport: a pipeline for portuguese nlp (short paper)," in *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[45] R. Al-Rfou, "Polyglot: A massive multilingual natural language processing pipeline," Ph.D. dissertation, State University of New York at Stony Brook, 2015.

[46] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *convolutional neural networks and incremental parsing*, vol. 7, 2017.

[47] L. Padró and E. Stanilovsky, “Freeling 3.0: Towards wider multilinguality,” in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA, May 2012.

[48] L. Màrquez and H. Rodríguez, “Part-of-speech tagging using decision trees,” in *Machine Learning: ECML-98*, C. Nédellec and C. Rouveirol, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 25–36.

[49] P. Qi, T. Dozat, Y. Zhang, and C. D. Manning, “Universal dependency parsing from scratch,” in *CoNLL Shared Task*, 2018.

[50] J. Baldridge, “The opennlp project,” *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*, p. 1, 2005.

[51] E. R. Fonseca and J. L. G. Rosa, “Mac-morpho revisited: Towards robust part-of-speech tagging,” in *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013. [Online]. Available: https://www.aclweb.org/anthology/W13-4811

[52] S. Buchholz and E. Marsi, “Conll-x shared task on multilingual dependency parsing,” in *Proceedings of the Tenth Conference on Computational Natural Language Learning*, ser. CoNLL-X ’06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 149–164. [Online]. Available: http://dl.acm.org/citation.cfm?id=1596276.1596305

[53] A. R. O. Pires, “Named entity extraction from portuguese web text,” Master’s thesis, Faculdade de Engenharia da Universidade do Porto, 2017.

[54] D. Jurafsky and J. H. Martin, “Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition (Third Edition draft),” 2019. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[55] ——, “Chapter 8: Part-of-speech tagging,” *Speech and Language Processing*, 2019.

[56] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 282–289.

[57] G. Tambouratzis, M. Troullinos, S. Sofianopoulos, and M. Vassiliou, “Accurate phrase alignment in a bilingual corpus for ebmt systems,” in *Proceedings of the 5th BUCC Workshop, held within the LREC2012 Conference*, vol. 26. Citeseer, 2012, pp. 104–111.

[58] F. Sánchez-Martínez and M. L. Forcada, "Inferring shallow-transfer machine translation rules from small parallel corpora," *Journal of Artificial Intelligence Research*, vol. 34, pp. 605–635, 2009.

[59] H. G. Oliveira, V. de Paiva, C. Freitas, A. Rademaker, L. Real, and A. Simões, "As wordnets do português," *Oslo Studies in Language*, vol. 7, no. 1, 2015.

[60] A. Farkiya, P. Saini, S. Sinha, and S. Desai, "Natural language processing using nltk and wordnet," *International Journal of Computer Science and Information Technologies*, vol. 6, 2015.

[61] N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, and S. Aluisio, "Portuguese word embeddings: Evaluating on word analogies and natural language tasks," *arXiv preprint arXiv:1708.06025*, 2017.

[62] A. M. Mara Moita, Helena Carmo, "O comportamento sintático das interrogativas-q na língua gestual portuguesa: Estudo preliminar." II Encontro sobre Morfossintaxe da Língua Gestual Portuguesa e outras línguas de sinais, 2018.

[63] N. Santana, "Aspeto verbal na lgp," *Exedra: Revista Científica*, no. 6, pp. 373–377, 2012.

[64] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[65] C.-H. Wu, H.-Y. Su, Y. Chiu, and C.-H. Lin, "Transfer-based statistical translation of taiwanese sign language using pcfg," *ACM Trans. Asian Lang. Inf. Process.*, vol. 6, 04 2007.

[66] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040

[67] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, vol. 200, no. 6. Cambridge, MA, 2006.

[68] B. Dorr, M. Snover, and N. Madnani, "Part 5: Machine translation evaluation," *Dostopljeno na: https://www. cs. cmu. edu/˜ alavie/papers/GALE-book-Ch5. pdf [8. 8. 2018]*, 2006.

[69] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, vol. 200, no. 6, 2006.

[70] C. Choupina, A. M. Barros Brito, and F. Bettencourt, “Particularidades da morfossintaxe das construções ditransitivas com o verbo ’dar’ na língua gestual portuguesa,” *Revista da Associação Portuguesa de Linguística*, pp. 117–147, 01 2016.

A

Descrição da coleção dourada

Tipo	Fonte	%
Revistas	Visão	30
	Exame Informática	29
Jornais	Observador	27
	Público	2
Livros		13

Tabela A.1: Composição da coleção dourada.

Classes	Frequência
Organização	30
Localização	22
Pessoa	8
Data	21

Tabela A.3: Frequência de cada classe de entidades nomeadas na coleção dourada.

Classes	**Sub-classes**	**Frequência**
Adjetivo (Adj)		179
Advérbio (Adv)	Negação (N)	15
	Normal (G)	73
Conjunção (Conj)	Coordenativa (C)	93
	Subordinativa (S)	58
Determinante (Det)	Artigo (Art)	468
	Indefinido (Ind)	24
	Possessivo (Poss)	14
	Demonstrativo (Dem)	35
	Interrogativo (Int)	2
Nome (Nom)	Comum (C)	699
	Próprio (P)	109
Numeral (Num)		77
Preposição (Prep)		525
Pronome (Pron)	Pessoal (Pes)	23
	Indefinido (Ind)	4
	Relativo (Rel)	43
	Demonstrativo (Dem)	16
Verbo (Verb)	Auxiliar (Aux)	82
	Indicativo (Ind)	142
	Condicional (Cond)	3
	Conjuntivo (Conj)	5
	Gerúndio (Ger)	13
	Particípio Passado (PP)	98
	Infinitivo (Inf)	74

Tabela A.2: Frequência de cada classe morfossintática na coleção dourada.

B

Resultados da avaliação de ferramentas de NLP

Ferramenta	Micro-Média	Macro-Média
NLTK (Bigramas)	0.87	0.69
NLTK (Perceptrão)	0.89	0.71
NLTK (Máxima Entropia)	0.93	0.74
NLPyPort	0.86	0.83
Polyglot	0.79	0.68
Spacy	0.90	0.47
FreeLing	0.90	0.58
TreeTagger	0.91	0.73
StanfordNLP	0.91	0.61
OpenNLP (Perceptrão)	0.93	0.77
OpenNLP (Máxima Entropia)	**0.94**	**0.91**

Tabela B.1: Micro- e Macro-Média relativos a F1 para os diferentes modelos na tarefa de análise morfossintática.

Classes		NLTK			NLPy	PG	SpaCy	TT	FL	StfNLP	ONLP	
		Bg	P	ME							ME	P
Adj		0.76	0.82	0.83	0.75	0.63	0.89	0.92	**0.96**	0.94	0.88	0.85
Adv	**G**	0.81	0.77	0.81	0.80	0.46	0.85	**0.91**	0.84	0.89	0.82	0.81
	N							**0.97**	**0.97**	**0.97**		
Conj	**C**	0.97	0.95	0.96	0.97	0.94	0.95	0.97	0.94	0.97	0.97	**0.98**
	S	0.56	0.72	0.60	0.60	0.56	0.68	0.87	**0.88**	0.74	0.70	0.70
Det	**Art**	0.95	0.93	0.96	0.97	-	0.97	0.94	**0.99**	0.98	0.96	0.97
Nom	**C**	0.82	0.91	0.93	0.81	0.93	0.95	0.94	**0.98**	**0.98**	0.94	0.94
	P	0.33	0.33	0.81	0.33	0.68	0.85	0.51	**0.96**	0.85	0.77	0.75
Num		0.84	0.95	**0.98**	0.83	0.72	0.96	**0.98**	0.95	0.97	0.94	0.97
Prep		0.95	0.95	0.96	0.97	0.85	0.96	0.97	**0.99**	0.96	0.96	0.96
Pron	**Pes**	0.97	0.90	0.90	**1.00**	-	0.83	0.98	0.84	**1.00**	0.91	0.94
Verb	**Aux**	0.88	0.94	0.97	0.88	0.00	**0.84**	0.00	0.00	0.73	0.96	0.97
	Ind						**0.88**	0.82	0.83	0.85		
	Cond						0.50	*		**0.80**		
	Conj					0.62	0.36	0.74	0.71	**0.92**		
	Ger	0.91	1.00	0.91	0.91		0.95	0.91	0.91	0.95	**1.00**	0.96
	PP	0.82	0.95	0.92	0.82		0.94	**0.96**	0.95	0.93	**0.96**	**0.96**
	Inf	0.89	0.93	0.95	0.89		0.91	0.92	0.94	0.95	**0.96**	0.95

Tabela B.2: Valores de F1 para as categorias comuns. O * indica que o Condicional é visto como Indicativo por estes sistemas

Classes		Polyglot	TreeTagger	FreeLing	StanfordNLP
Det	**Ind**	0.79	0.96	**1.00**	0.91
	Poss		1.00	1.00	1.00
	Dem		0.89	1.00	0.97
	Int		0.00	0.00	1.00
Pron	**Ind**	0.70	0.46	**0.89**	0.75
	Rel		0.93	**0.94**	0.94
	Dem		0.79	0.86	**0.90**

Tabela B.3: Valores de F1 para as ferramentas que têm as categorias **Det** e **Pron** mais finas

Classes	NLTK			NLPy	SpaCy	OpenNLP	
	Bigramas	Perc.	ME			ME	Perc.
Pron-det	0.83	0.79	0.86	0.88	0.81	**0.89**	0.88
Pron-indp	0.80	0.88	0.82	0.79	**0.91**	0.86	0.88

Tabela B.4: Valores de F1 para as etiquetas **Pron-det** e **Pron-indp**. A primeira contém os determinantes, pronomes demonstrativos, pronomes interrogativos, pronomes possessivos e pronomes relativos; a segunda os pronomes indefinidos e outros pronomes de outras categorias que expressam imprecisão.

Ferramenta	Micro-Média F1	Macro-Média F1
NLTK (Árvore de Decisão)	0.97	0.47
NLTK (Naïve Bayes)	0.92	0.18
NLTK (Máxima Entropia)	0.97	0.35
Polyglot	0.98	0.76
FreeLing	**0.99**	0.77
StanfordNLP	0.98	**0.78**
OpenNLP	0.97	0.46

Tabela B.5: Micro- e Macro-Média relativos a F1 para os diferentes modelos na tarefa de reconhecimento de entidades nomeadas.

			NLTK				
Classes	**FreeLing**	**Polyglot**	**AD**	**NB**	**EM**	**OpenNLP**	**StanfordNLP**
Data	–	–	0.78	0.11	0.74	0.76	**0.92**
Localizacao	**0.92**	0.73	0.51	0.06	0.17	0.00	0.34
Organizacao	**0.84**	0.61	0.62	0.34	0.58	0.70	0.75
Pessoa	0.67	0.70	0.42	0.00	0.35	0.32	**0.88**

Tabela B.6: Valores de F1 de cada ferramenta, tendo em conta as entidades nomeadas da coleção dourada.

C

Algoritmo do alinhamento de palavras e glosas

Algoritmo C.1: Algoritmo de alinhamento dos lemas e glosas.

```
begin
    if lema == gesto then
        alinhar(lema, gesto);
    else
        if wup_palmer(sinónimo_lema, sinónimo_gesto) >= 0.90 then
            alinhar(lema, gesto);
        else
            if Jaro_Winkler(sinónimo_lema, sinónimo_gesto) >= 0.80 then
                alinhar(lema, gesto);
                else
                    if word_embeddings(lema, gesto) > 0.3 then
                        alinhar(lema, gesto);
                    else
                        não_alinhar(lema,gesto)
```

D

Dicionário bilingue

Amostra de entradas do dicionário bilingue.

Palavra	**Lema da palavra**	**Gesto**
houve grande	haver grande	ter-muito
desenvolvimento	desenvolvimento	desenvolvimento
político	político	político
religioso	religioso	igreja
foi muito	ir muito	ter-muito
estado	estado	estado
religião	religião	igreja
investiu	investir	investir
pouco	pouco	pouco
financeiramente	financeiro	dinheiro
isto	isto	ele
bom	bom	bom
aproveitou	aproveitar	aproveitar
mudança	mudança	mudar
conflitos	conflito	conflito

E

Corpus de desenvolvimento

A Maria foi à China.
A Maria não come carne.
A Marta e o José foram às compras.
O meu amigo João gosta de aranhas!
O João gosta da minha amiga Maria.
A leoazinha do jardim zoológico já cresceu.
Vamos comprar aquela mesinha!
Quem é que molhou o elefante?
Quem está a molhar o elefante?
Quem é que o gato morde?
Quem o elefante molhou?
Quem o gato arranha?
Quem está a morder a vaca?
Quem espantou o gato?
Qual dos bolos preferes?
Quantos anos tens?
Que idade tens?
Quanto custa?
Onde está a Maria?
Ontem, o Manuel José cozinhou um bolo muito bom!
Amanhã irei comprar dois livros à livraria.
Queres vir comigo?
Quando a Maria comeu?
A Matilde foi fazer compras à China.
Na próxima segunda-feira irei telefonar à minha mãe.
As minhas irmãs são gémeas.
Eles disseram que em 2008 houve uma grande crise económica!
Ele disse que tu gostavas de nadar.
O acidente aconteceu ontem de madrugada.
A minha mãe estava chateada quando eu falei com ela.
Ela entrou no quarto triste!
Quando é que tens consulta?
Tem dificuldades respiratórias?
A dor é persistente?
Como posso ajudá-lo?
Desculpe, como podemos chegar à estação?
A minha mãe trabalha como rececionista num escritório.
Tu compraste-me 4 relógios.
Tu compraste-me quatro relógios.
O gatinho dela é tão fofo!
Eles disseram que a aula ia ser por Zoom.
Foi muito importante o que disseste ontem!
Qual das éguas é tua?
O gato da Rute é muito meigo e lindo!
Nós, ontem fomos a Évora.
Carolina, queres vir ao cinema?
Qual é o teu filme favorito.
Aquela árvore está cheia de abelhões!
A chave está no barracão.
Que caixinha tão bonita!
Dói-te muito?
Onde te dói?
Tiveste febre?
A temperatura ultrapassou os 38.5 graus?
Querem chocolate?
As nossas vizinhas já voltaram?
Esse livro é meu!
Amanhã vai chover mas no fim de semana não!
Bem e como vais tu?
A Maria que teve um filho há pouco tempo adora morangos.
Quando fazes anos?
Quando é que fazes anos?
A Maria foi mordida por um cão?
Qual foi o filme que viste ontem com o Jorge?
As éguas da minha tia saltam muito alto!
Qual é o nome desta estrada?
Qual é o teu nome?
O meu nome é Joana Silva!
Vou chamar o médico.
Chamas-me às 6 horas da manhã?
Chamas um táxi para mim, por favor?
Ontem choveu!
Ontem tive um furo no pneu.
A queda de água é linda.
Estava frio hoje de manhã.
O seu quarto é no segundo piso, à direita.
Qual é a temperatura?
Qual é o teu número de telemóvel?

F

Traduções do sistema baseline

Frase em português	Frase em português gestuado
Bom dia.	BOM DIA
A menina de óculos vê a flor.	MENINA ÓCULOS VER FLOR
O aluno gosta de animais.	ALUNO GOSTAR ANIMAIS
O segurança também quer respeito.	SEGURANÇA TAMBÉM QUERER RESPEITO
O jovem quer comida.	JOVEM QUERER COMER
Eles gostam de massa.	ELES GOSTAR MASSA
O neto come a formiga.	NETO COMER FORMIGA
O diretor pede dinheiro.	DIRETOR PEDIR DINHEIRO
A criança é inteligente.	CRIANÇA SER INTELIGENTE
O estado tem o poder	ESTADO TER PODER
O pai zangou-se com o filho.	PAI ZANGAR SE FILHO
Ele é espanhol.	ELE SER ESPANHOL
O papa é bom ouvinte	PAPA SER BOM OUVINTE
O meu namorado tem olhos verdes.	MEU NAMORADO TER OLHOS VERDES
O médico ouve uma história.	MÉDICO OUVIR HISTÓRIA
Eu vejo o meu filho.	EU VER MEU FILHO
Ela tem uma casa azul.	ELA TER CASA AZUL
A mulher cega é simpática.	MULHER CEGA SER SIMPÁTICA
Eu pergunto-te.	EU PERGUNTAR TE
Ali está pouco sol.	ALI ESTAR POUCO SOL
A menina de óculos não vê a flor.	MENINA ÓCULOS NÃO VER FLOR
O aluno não gosta de animais.	ALUNO NÃO GOSTAR ANIMAIS
O segurança não quer respeito.	SEGURANÇA NÃO QUERER RESPEITO
O jovem não quer comida.	JOVEM NÃO QUERER COMER
Eles não gostam de massa.	ELES NÃO GOSTAR MASSA
O neto não come a formiga.	NETO NÃO COMER FORMIGA
O diretor não pede dinheiro.	DIRETOR NÃO PEDIR DINHEIRO
A criança não é inteligente.	CRIANÇA NÃO SER INTELIGENTE
O estado não tem o poder	ESTADO NÃO TER PODER
O pai não se zangou com o filho.	PAI NÃO SE ZANGAR FILHO
Ele não é espanhol.	ELE NÃO SER ESPANHOL
O papa não é bom ouvinte	PAPA NÃO SER BOM OUVINTE
O meu namorado não tem olhos verdes.	MEU NAMORADO NÃO TER OLHOS VERDES
O médico não ouve uma história.	MÉDICO NÃO OUVIR HISTÓRIA
Eu não vejo o meu filho.	EU NÃO VER MEU FILHO
Ela não tem uma casa azul.	ELA NÃO TER CASA AZUL
A mulher cega não é simpática.	MULHER CEGA NÃO SER SIMPÁTICA
Eu não te pergunto.	EU NÃO TE PERGUNTAR
Ali não está pouco sol.	ALI NÃO ESTAR POUCO SOL
A menina de óculos vê a flor?	MENINA ÓCULOS VER FLOR
O aluno gosta de animais?	ALUNO GOSTAR ANIMAIS
O segurança quer respeito?	SEGURANÇA QUER RESPEITO
O jovem quer comida?	JOVEM QUERER COMER
Eles gostam de massa?	ELES GOSTAR MASSA
O neto come a formiga?	NETO COMER FORMIGA
O diretor pede dinheiro?	DIRETOR PEDIR DINHEIRO
A criança é inteligente?	CRIANÇA SER INTELIGENTE
O estado tem o poder?	ESTADO TER PODER
O pai zangou-se com o filho?	PAI ZANGAR SE FILHO
Ele é espanhol?	ELE SER ESPANHOL
O papa é bom ouvinte?	PAPA SER BOM OUVINTE
O meu namorado tem olhos verdes?	MEU NAMORADO TER OLHOS VERDES
O médico ouve uma história?	MÉDICO OUVIR HISTÓRIA
Eu vejo o meu filho?	EU VER MEU FILHO
Ela tem uma casa azul?	ELA TER CASA AZUL
A mulher cega é simpática?	MULHER CEGA SER SIMPÁTICA
Eu pergunto-te?	EU PERGUNTAR TE
Ali está pouco sol?	ALI ESTAR POUCO SOL

G

Valores de TER das traduções do PE2LGP

Traduções do sistema PE2LGP	**TER**
BOM-DIA	0
MULHER MENINO ÓCULOS FLOR VER	0.2
ALUNO ANIMAIS GOSTAR	0.67
SEGURANÇA TAMBÉM RESPEITO QUERER	0.25
JOVEM COMER QUERER	0.67
ELES MASSA GOSTAR	0.33
NETO FORMIGA COMER	0.33
DIRETOR DINHEIRO PEDIR	0.33
CRIANÇA INTELIGENTE	0
ESTADO PODER TER	0.5
PASSADO PAI FILHO ZANGAR	1
ELE ESPANHOL	1
PAPA BOM OUVINTE	0.33
NAMORADO MEU VERDES OLHOS TER	0.75
MÉDICO HISTÓRIA OUVIR	0.33
EU FILHO MEU VER	0.25
ELA CASA AZUL TER	0.75
MULHER CEGA SIMPÁTICA	0
EU TU PERGUNTAR	0.33
ALI ALI POUCO SOL	0.67
ÓCULOS MULHER MENINO FLOR VER {NÃO}(headshake)	0.6
ALUNO ANIMAIS GOSTAR {NÃO}(headshake)	0.5
SEGURANÇA RESPEITO QUERER {NÃO}(headshake)	0.67
JOVEM COMER QUERER {NÃO}(headshake)	1
ELES MASSA GOSTAR {NÃO}(headshake)	0.5
NETO FORMIGA COMER {NÃO}(headshake)	0.5
DIRETOR DINHEIRO PEDIR {NÃO}(headshake)	0.5
CRIANÇA INTELIGENTE {NÃO}(headshake)	0
ESTADO PODER TER {NÃO}(headshake)	0.67
PASSADO PAI FILHO ZANGAR {NÃO}(headshake)	1
ELE ESPANHOL {NÃO}(headshake)	1
PAPA BOM OUVINTE {NÃO}(headshake)	0.25
NAMORADO MEU OLHOS VERDES TER {NÃO}(headshake)	0.2
MÉDICO HISTÓRIA OUVIR {NÃO}(headshake)	0.5
EU FILHO MEU VER {NÃO}(headshake)	0.4
ELA CASA AZUL TER {NÃO}(headshake)	0.6
MULHER CEGA SIMPÁTICA {NÃO}(headshake)	0.25
EU TU PERGUNTAR {NÃO}(headshake)	0.5
ALI POUCO SOL {NÃO}(headshake) {NÃO}(headshake)	0.5
{MULHER MENINO ÓCULOS FLOR VER}(q)	0
{ALUNO ANIMAIS GOSTAR}(q)	0.33
{SEGURANÇA RESPEITO QUER}(q)	0.33
{JOVEM COMER QUERER}(q)	0.33
{ELES MASSA GOSTAR}(q)	0
{NETO FORMIGA COMER}(q)	0
{DIRETOR DINHEIRO PEDIR}(q)	0
{CRIANÇA INTELIGENTE}(q)	0
{ESTADO PODER TER}(q)	0.33
{PASSADO PAI FILHO ZANGAR}(q)	0.67
{ELE ESPANHOL}(q)	1
{PAPA OUVINTE BOM}(q)	0
{NAMORADO MEU OLHOS VERDES TER}(q)	0.5
{MÉDICO HISTÓRIA OUVIR}(q)	0
{EU FILHO MEU VER}(q)	0
{ELA AZUL CASA TER}(q)	0.75
{MULHER CEGA SIMPÁTICA}(q)	0
{EU TU PERGUNTAR}(q)	0
{ALI SOL POUCO}(q)	0

H

Questionário

Avaliação do tradutor de português europeu para língua gestual portuguesa

No âmbito do Mestrado em Engenharia Informática e de Computadores do Instituto Superior Técnico desenvolveu-se um tradutor automático de texto em português para língua gestual portuguesa (LGP) baseado em regras extraídas de um corpus paralelo de português e LGP. O resultado do tradutor é uma sequência de glosas com marcadores que identificam expressões faciais gramaticais e outros elementos.

Para avaliar a tradução do sistema desenvolvido peço a sua colaboração neste questionário. O questionário consiste na análise e classificação da tradução automática (sequência de glosas) de 13 frases em português. Ao longo do mesmo poderá ser pedido para justificar a sua escolha. Mas atenção! O objetivo não é avaliar as suas respostas mas sim o output do sistema de tradução!
O questionário não demorará mais do que 30 minutos e as suas respostas permanecerão anónimas, sendo apenas usadas para análise.

Desde já, agradeço a sua colaboração.

Como classifica o seu nível de língua gestual portuguesa?

- Nativo
- Fluente
- Intermédio
- Elementar
- Principiante

Frase 1

Para relembrar

Antes de começar, é preciso que se familiarize com as notações presentes nas sequências de glosas para identificar alguns fenómenos linguísticos da LGP:

> As expressões faciais são identificadas dentro de parênteses, **()**. Por exemplo, a expressão facial **(headshake)** corresponde à expressão facial abanar a cabeça de um lado para o outro (na negação de verbos). Por sua vez, **(q)** corresponde à expressão facial que marca as frases interrogativas (levantamento do queixo com inclinação da cabeça para trás e franzir das sobrancelhas).
> A duração das expressões é identificada por chavetas **{}**, em que o início da expressão facial é marcado por **{** e o fim por **}**.

Tarefa: Indique qual a tradução em português da seguinte sequência de glosas.

Nota: existem várias traduções possíveis.

SEGURANÇA QUERER TAMBÉM RESPEITO

Tarefa: Responda às seguintes questões com base na sequência de glosas anterior e uma das possíveis traduções em português dadas em seguida.

Para relembrar

Antes de começar, é preciso que se familiarize com as notações presentes nas sequências de glosas para identificar alguns fenómenos linguísticos da LGP:

> As expressões faciais são identificadas dentro de parênteses, **()**. Por exemplo, a expressão facial **(headshake)** corresponde à expressão facial abanar a cabeça de um lado para o outro (na negação de verbos). Por sua vez, **(q)** corresponde à expressão facial que marca as frases interrogativas (levantamento do queixo com inclinação da cabeça para trás e franzir das sobrancelhas).
> A duração das expressões é identificada por chavetas **{}**, em que o início da expressão facial é marcado por **{** e o fim por **}**.

Sequência de glosas (tradução): SEGURANÇA QUERER TAMBÉM RESPEITO
Frase em português: O segurança também quer respeito.

Classifique a qualidade da tradução em LGP:

Pobre - A tradução está incorreta (o significado da tradução está incorreto).
Justo - O significado da tradução é o correto mas a gramática falha em alguns aspetos.
Bom - O significado da tradução e a gramática estão corretos.

Tradução

○ Pobre ○ Justo ○ Bom

Sequência de glosas (tradução): SEGURANÇA QUERER TAMBÉM RESPEITO
Frase em português: O segurança também quer respeito.

Classifique a tradução para LGP com base nas seguintes características:

Ordem frásica vs Ordem das glosas
Ordem frásica - Ordem do sujeito, verbo e objeto.
Ordem das glosas - Ordem de constituintes como determinantes possessivos, pronomes interrogativos, etc.

Ordem frásica

Incorreto Parcialmente (in)correto Correto Não se aplica

Ordem das glosas

Incorreto Parcialmente (in)correto Correto Não se aplica

Léxico

Incorreto Parcialmente (in)correto Correto Não se aplica

Expressão facial

Incorreto Parcialmente (in)correto Correto Não se aplica

Duração da expressão facial

Incorreto Parcialmente (in)correto Correto Não se aplica

I

Frases da avaliação manual

Frase em português	**Tradução pelo sistema PE2LGP**
O segurança também quer respeito.	SEGURANÇA QUERER TAMBÉM RESPEITO
O pai zangou-se com o filho.	PASSADO PAI ZANGAR FILHO
Ele é espanhol.	ELE ESPANHOL
Ela tem uma casa azul.	ELA TER CASA AZUL
Eles não gostam de massa.	ELES GOSTAR MASSA {NÃO}(headshake)
O meu namorado não tem olhos verdes.	NAMORADO MEU TER OLHOS VERDES {NÃO}(headshake)
Eu não vejo o meu filho.	EU VER FILHO MEU {NÃO}(headshake)
A mulher cega não é simpática.	MULHER CEGA SIMPÁTICA {NÃO}(headshake)
Eu não te pergunto.	EU PERGUNTAR TU {NÃO}(headshake)
O aluno gosta de animais?	{ALUNO ANIMAIS GOSTAR}(q)
O estado tem o poder?	{ESTADO PODER TER}(q)