Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 13, Número 1 (2021)

ISSN: 1647-0818

linguaMATICA

Volume 13, Número 1 – 2021

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Artigos de Investigação

Editorial

Embora não sejamos supersticiosos, não podemos deixar de fazer a relação entre o décimo terceiro ano de vida da Linguamática, e a súbita descida no índice SciMago, para o terceiro quartil. Esperamos que nos próximos anos consigamos recuperar, colocando a Linguamática entre as melhores revistas da área.

Esta edição conta apenas com dois artigos: um sobre tradução automática da língua portuguesa em língua guestual portuguesa, e um outro de estudo de um corpus *Espanhol–Inglês–Chinês.*

Como sempre, o nosso muito obrigado a todos os revisores e autores que continuam a acreditar neste nosso/vosso projeto.

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís,
Universidade de Vigo

Alberto Simões,
Instituto Politécnico do Cávado e Ave

Aline Villavicencio,
Universidade Federal do
Rio Grande do Sul

Álvaro Iriarte Sanroman,
Universidade do Minho

Anselmo Peñas,
Universidad Nacional de
Educación a Distancia

Antón Santamarina,
Universidade de Santiago de
Compostela

Antoni Oliver González,
Universitat Oberta de Catalunya

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arkaitz Zubiaga,
Dublin Institute of Technology

Bruno Martins,
Instituto Superior Técnico

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Fernando Batista,
Instituto Universitário de Lisboa

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Université de Caen Basse-Normandie

Gerardo Sierra,
Universidad Nacional
Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Irene Castellón Masalles,
Universitat de Barcelona

Iria da Cunha,
Universidad Nacional de
Educación a Distancia

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Juan-Manuel Torres-Moreno,
Université d´Avignon et
des Pays du Vaucluse

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Luís Morgado da Costa,
Nanyang Technological University

Manex Agirrezabal,
University of Copenhagen

Marcos Garcia,
Universidade da Corunha

María Inés Torres,
Euskal Herriko Unibertsitatea

Mário Rodrigues,
Universidade de Aveiro

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Miguel Solla Portela,
Universidade de Vigo

Mikel Forcada,
Universitat d'Alacant

Pablo Gamallo Otero,
Universidade de Santiago de
Compostela

Patrícia Cunha França,
Universidade do Minho

Patricia Martin Rodilla
Universidade de Santiago de
Compostela

Ricardo Rodrigues
Instituto Politécnico de Coimbra

Rui Pedro Marques,
Universidade de Lisboa

Susana Afonso Cavadas,
University of Exeter

Xavier Gómez Guinovart,
Universidade de Vigo

Artigos de Investigação

Proposta recebida em Setembro 2020 e aceite para publicação em Março 2021.

PE2LGP: tradutor de português europeu para língua gestual portuguesa em glosas

PE2LGP: translating European Portuguese into Portuguese Sign Language glosses

Matilde Gonçalves
Instituto Superior Técnico
Universidade de Lisboa
INESC-ID Lisboa

Luísa Coheur
Instituto Superior Técnico
Universidade de Lisboa
INESC-ID Lisboa

Hugo Nicolau
Instituto Superior Técnico
Universidade de Lisboa
INESC-ID Lisboa

Ana Mineiro
Instituto de Ciências da Saúde
Universidade Católica Portuguesa
Centro de Investigação Interdisciplinar em Saúde

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 e que as traduções estejam semanticamente corretas. Trabalhos desenvolvidos anteriormente utilizam exclusivamente regras de tradução manuais, sendo muito limitados na quantidade de fenómenos gramaticais abrangidos, produzindo pouco mais que português gestuado.

Neste artigo, apresenta-se o primeiro sistema de tradução de português para a língua gestual portuguesa, o PE2LGP, que, para além de regras manuais, se baseia em regras de tradução construídas automaticamente a partir de um corpus de referência. Dada uma frase em português, o sistema devolve uma sequência de glosas com marcadores que identificam expressões faciais, palavras soletradas, entre outras. Uma avaliação automática e uma avaliação manual são apresentadas, indicando os resultados melhorias na qualidade da tradução de frases simples e pequenas em comparação ao sistema *baseline* (português gestuado).

Este é, também, o primeiro trabalho que lida com as expressões faciais gramaticais que marcam as frases interrogativas e negativas.

Palavras chave

português europeu, língua gestual portuguesa, tradução automática, corpus anotado, glosa, processamento da linguagem natural

Abstract

As the Portuguese language, the 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 (signed Portuguese), as it should ensure that the resulting signs satisfy the grammar of the Portuguese sign language and that the translations are semantically correct. Previous works use exclusively manual translation rules and are very limited in the amount of grammatical phenomena covered, producing merely signed Portuguese.

This paper presents the first translator from Portuguese to the Portuguese sign language, based on manual rules, but also in translation rules automatically built from a reference corpus. Given a sentence in Portuguese, the system returns a sequence of glosses with markers that identify facial expressions, spelled words, among others. The paper reports both a manual and automatic evaluation. Results show improvements in the translation quality of simple and short sentences compared to the baseline system ("signed Portuguese").

Moreover, this is the first study that deals with grammatical facial expressions, which mark interrogative and negative sentences.

Keywords

European Portuguese, Portuguese sign language, automatic translation, annotated corpus, gloss, natural language processing

DOI: 10.21814/lm.13.1.338

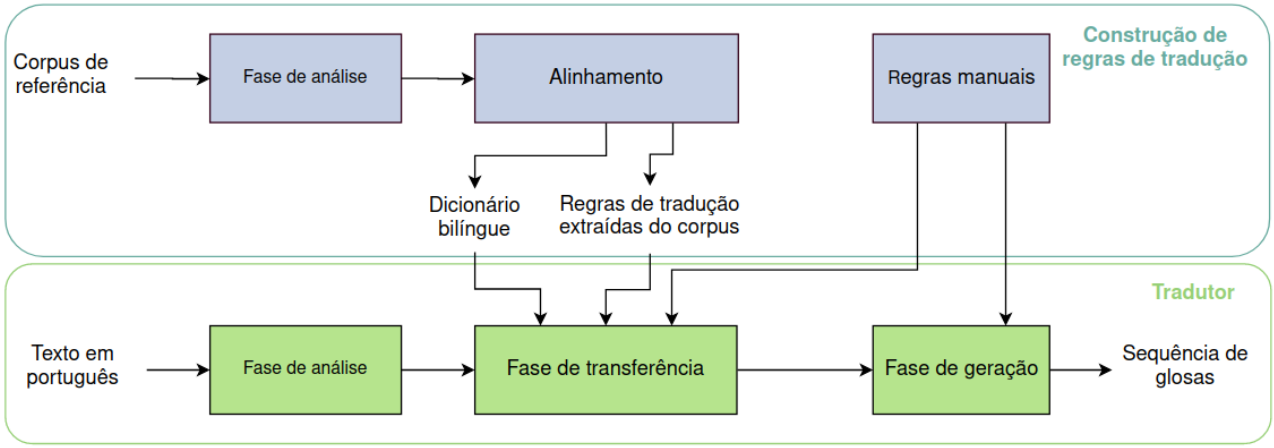

Figura 1: Arquitetura do sistema de tradução PE2LGP

1. Introdução

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 a comunidade surda, e também para fins de aprendizagem da LGP. No entanto, a LGP apresenta várias diferenças gramaticais em relação à língua portuguesa. Assim, um tradutor que não queira apenas gerar "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 especificidades da LGP. Apesar de existirem alguns estudos linguísticos sobre esta (Amaral et al., 1994; Bettencourt, 2015; Choupina, 2017; Choupina et al., 2017), bem como gestuários/dicionários (Ferreira, 1997; Mesquita & Silva, 2009; Baltazar, 2010), não existe ainda uma gramática oficial, nem sequer consenso sobre variados fenómenos linguísticos. Por exemplo, alguns autores consideram que a estrutura base das frases é Sujeito–Verbo–Objeto (SVO), outros Sujeito–Objeto–Verbo (SOV). Talvez por isso os poucos trabalhos computacionais ligados à tradução para LGP (Almeida et al., 2015b; Escudeiro et al., 2015; dos Santos, 2016; Ferreira, 2016; Gaspar, 2015) focam pouco a componente linguística, baseando-se em pequenos conjuntos de regras manuais e excluindo expressões faciais, resultando em pouco mais 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.

Neste trabalho, contribuímos com um tradutor para LGP, doravante PE2LGP, em que as frases traduzidas para LGP são representadas por sequências de glosas, com marcadores que identificam as expressões faciais e palavras soletradas. O PE2LGP apoia-se em regras de tradução e num dicionário bilingue criados automaticamente a partir do corpus referido; adicionalmente um conjunto de regras manuais pode ser adicionado. A Figura 1 ilustra a arquitectura do PE2LGP, que segue a estrutura tradicional dos sistemas de tradução baseados na transferência gramatical. O PE2LGP começa por extrair informação do corpus e enriquecê-la com informação linguística. Seguidamente, procede-se ao alinhamento entre as palavras e os gestos do corpus. Deste alinhamento são extraídas as regras de tradução e um dicionário bilingue de português e LGP. Quando é dada ao sistema uma (ou mais) frase(s) em português, depois de um pré-processamento linguístico, entra em acção o módulo de tradução, que, com base nos recursos anteriormente criados, faz a sua tradução para LGP. Para além das regras de tradução extraídas automaticamente do corpus e do dicionário bilingue, 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, que as regras de tradução não cobrem, tais como as expressões faciais.

Neste artigo, apresentamos ainda duas avaliações do PE2LGP, uma automática, com base num corpus de teste construído por especialistas e outra manual, em que falantes de LGP avaliam a qualidade das traduções. De notar que

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

o PE2LGP permite ainda que se gerem frases segundo a ordem SOV ou SVO. Estas duas hipóteses foram também avaliadas.

A principal contribuição deste trabalho é um tradutor entre português europeu e LGP, que se alimenta de um corpus de referência para criar regras de tradução e um dicionário bilingue (podendo, portanto, crescer com o corpus). No entanto, contribuímos ainda com:

- um método de alinhamento, baseado em *string matching* e semelhança semântica, tirando partido da *OpenWordNet-PT*[2] e de *word embeddings*;
- um conjunto de regras manuais;
- um módulo que recolhe informações estatísticas das regras extraídas do corpus.

De acordo com o nosso conhecimento, 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. De notar que, se uma frase dada não for apanhada pelas regras do tradutor, o processo continua, resultando em "português gestuado". Todos os recursos desenvolvidos neste trabalho são *open-source*[3].

Este documento está organizado em mais cinco secções: na Secção 2 são apontados alguns aspetos da gramática da língua gestual portuguesa. A revisão da literatura encontra-se na Secção 3. Nas Secções 4 e 5 descreve-se o PE2LGP. A metodologia de avaliação e os resultados são apresentados na Secção 6. Por fim, na Secção 7 resume-se as principais conclusões e o trabalho futuro.

2. Sobre a língua gestual portuguesa

Os primeiros estudos sobre a LGP surgiram na década de 90, não existindo ainda uma gramática oficial. Nesta secção descrevem-se alguns aspetos gramaticais da LGP. De notar que alguns fenómenos linguísticos frequentes não foram ainda estudados em LGP. É o caso de estruturas como a subordinação.

2.1. Estrutura frásica canónica

Como dito anteriormente, não existe ainda consenso sobre a ordem frásica base da LGP. Alguns autores defendem que a estrutura predominante é SOV (Rodrigues, 2018). No entanto, o estudo realizado por Bettencourt (2015), exactamente sobre a ordem canónica das frases em LGP, concluiu que, para frases com verbos transitivos não locativos e para frases declarativas, a ordem frásica base é igual à da língua portuguesa, isto é, SVO.

2.2. Tipos de frases

O tipo de frase, se é interrogativa ou negativa, influencia a ordem dos seus constituintes. De acordo com Bettencourt (2015), as frases interrogativas totais são marcadas pelo uso de advérbios e pronomes interrogativos, no final de uma frase em LGP, acompanhados pela expressão facial interrogativa.

2.3. Género feminino

Assumimos que a LGP é uma língua cuja marcação de género é apenas usada para explicitar o sexo de seres animados. Na verdade, investigadores como Choupina (2017) defendem que a LGP é uma língua sem sistema de género linguístico e sem sistema de número formal, ao contrário daquilo que é preconizado em (Amaral et al., 1994). No contexto deste artigo não nos propomos discutir esta questão e assumiremos uma posição operativa na senda de Amaral et al. (1994). Neste sentido, na LGP, essa marcação 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 (Bettencourt, 2015). Assim, o gesto LEÃO, como é um substantivo masculino, é representado apenas pelo gesto LEÃO, enquanto que LEOA é composto por MULHER + LEÃO. No entanto, existem situações em que não há marcação do género em nomes por existirem gestos para cada género associado ao nome. Por exemplo, os gestos para os nomes *galo* e *galinha* têm gestos próprios (Nascimento & Correia, 2011).

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. Assim, LEOAZINHA é composto pelos gestos MULHER + LEÃO + PEQUENO (com expressão facial).

[2] https://github.com/own-pt/openWordnet-PT
[3] https://github.com/mattgoncalves/PE2LGP

2.5. Plural

Existem quatro formas para marcar o plural (Bettencourt, 2015):

1. repetição do gesto usando a mão dominante. Por exemplo, para a palavra *árvores* é repetido o gesto ÁRVORE.
2. o gesto é produzido e repetido com as duas mãos (redobro). Um exemplo comum, é o caso do plural de *pessoa*, com as duas mãos repete-se o gesto PESSOA simultaneamente.
3. adição de um numeral que normalmente procede o substantivo. Por exemplo, *cinco livros* corresponde à sequência dos gestos LIVRO + CINCO.
4. adição de um advérbio de quantidade. Por exemplo, *muitos livros* corresponde a LIVRO + MUITO.

Os casos em que o plural é marcado com repetição ou redobro dos gestos não foram implementados neste tradutor.

2.6. Determinantes possessivos

Em LGP, determinantes possessivos (*meu*, *teu*, etc.) procedem o substantivo (Gaspar, 2015; Bettencourt, 2015). Por exemplo, *o teu irmão* originará a sequência de gestos: IRMÃO + TEU.

2.7. Tempos verbais

A marcação dos tempos verbais passado e futuro realiza-se de três formas (Nascimento & Correia, 2011):

- 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.

2.8. Verbos de concordância

Nas línguas gestuais existem gestos 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 (Choupina et al., 2016). Por exemplo, a produção em LGP da frase *Eu dou-te.* é apenas um gesto com posição inicial no EU (na pessoa que está a gestuar) e posição final no TU.

2.9. Negação

De acordo com Carmo et al. (2017), existem dois tipos de negação em LGP: a negação regular e a negação irregular. A primeira pode ser realizada pela adição de marcadores de negação manuais (por exemplo, a adição do gesto manual NÃO ou do gesto NADA depois do verbo), pela adição de gestos não manuais, como o marcador de negação *headshake* (abanar a cabeça de um lado para o outro repetidamente) ou pela 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).

2.10. 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.

2.11. Preposições

As preposições não são representadas em LGP isoladamente (Sousa, 2012); algumas são incorporadas no movimento dos gestos para identificar, por exemplo, os locais inicial e final do objeto que está em movimento (Bettencourt, 2015).

2.12. Conjunções coordenadas

De acordo com o estudo preliminar sobre conexões interfrásicas e frásicas (Martins & Mata, 2017), 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.

2.13. Classificadores

De acordo com Carmo (2016), 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 pode ser encontrada em (Carmo, 2016).

Dos fenómenos aqui descritos, além da marcação do plural por redobro ou por repetição de um gesto, os classificadores, a negação incorporada, os verbos de concordância e as preposições foram deixados para trabalho futuro.

3. Trabalho relacionado

A tradução para uma língua gestual pode ser feita com base em *corpora* e/ou regras manuais (Mohamed Amine, 2012). Caso exista uma quantidade razoável de textos alinhados entre a língua fonte e a língua gestual alvo, podem ser criados modelos computacionais com base nestes dados. Exemplos destes trabalhos são os sistemas de tradução para a língua gestual americana, apresentado em (Othman & Jemni, 2011) e para a alemã, descrito em (Bungeroth & Ney, 2004).

Como previamente referido, encontra-se em desenvolvimento, pela Universidade Católica Portuguesa, o primeiro corpus linguístico de referência da LGP, no qual as unidades lexicais são transcritas usando glosas e são anotadas informações gramaticais (classes gramaticais e análise sintática). Este corpus pode ser a fonte de informação de um modelo de tradução automática estatístico; neste trabalho tiramos partido deste corpus para extrair um conjunto de regras de tradução, às quais acrescentamos um conjunto de regras manuais.

Vários sistemas de tradução automática para língua gestual, baseados em regras, têm sido propostos nos últimos anos. Seguem-se alguns exemplos.

O projecto ATLASLang (Brour & Benabbou, 2019), um sistema híbrido de tradução de texto árabe (os autores não explicitam a variante do árabe) em língua gestual árabe, baseado em regras e em exemplos de frases (e das suas traduções) definidas num corpus bilingue. Se a frase existir nesse corpus, então é diretamente traduzida, caso contrário, a frase é processada e aplicam-se regras manuais. Em TEAM (Zhao et al., 2000), um protótipo de um sistema de tradução de texto inglês para língua gestual americana, as regras de tradução são definidas usando *tree-adjoining grammars* (Shieber & Schabes, 1990), resolvendo divergências linguísticas como a ordem das palavras nas frases. Referimos ainda o VLibras (Araújo et al., 2014), um sistema de tradução automática em tempo real de conteúdos digitais em português do Brasil para Língua Brasileira de Sinais (LIBRAS) através do processamento das legendas dos conteúdos multimédia. A tradução é baseada num pequeno conjunto de regras e os gestos são produzidos por um avatar 3D. Uma versão melhorada da componente de tradução deste sistema tendo em conta fenómenos sintáticos e semânticos da LIBRAS foi proposta por Lima et al. (2015). Outros exemplos, são os trabalhos de tradução de espanhol para Língua Gestual Espanhola (Lengua de Signos Española – LSE) (San-Segundo et al., 2006; Porta et al., 2014), que faz referência a um outro sistema de tradução espanhola para LSE baseado na plataforma Apertium (Forcada et al., 2011); tradução para língua gestual ucraniana em telemóveis (Davydov & Lozynska, 2017), e tradução de texto árabe para língua gestual árabe (Luqman & Mahmoud, 2018). A maioria usa a abordagem baseada na transferência gramatical (transferência sintática, lexical e semântica) através de regras de tradução criadas por linguistas.

Destaca-se o trabalho desenvolvido por Su & Wu (2009). 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 bilingues. As palavras e os gestos são alinhados usando uma medida de semelhança, em vez de métodos probabilísticos. Para a avaliação, foram traduzidas 50 frases retiradas de livros escolares chineses pelo presente sistema e o sistema baseline descrito em (Chiu et al., 2007). Os resultados mostram que o procedimento proposto pelos autores supera o sistema baseline, usando o corpus referido, principalmente em frases extensas. A estratégia implementada para o alinhamento de palavras e gestos neste trabalho foi uma fonte de inspiração para o nosso, 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.

Quanto à LGP, existem alguns protótipos computacionais, desenvolvidos recentemente, com objectivos distintos. Por exemplo, o trabalho de Bento (2013) foca-se em como levar um avatar a produzir gestos com base em gestos pro-

duzidos por humanos; em Gameiro et al. (2014) é proposto um sistema que visa o ensino de LGP. O "Virtual Sign Translator" (Escudeiro et al., 2013, 2015) contribui com um tradutor entre português e LGP, sendo também usado num jogo de ensino de LGP (Escudeiro et al., 2014). Almeida et al. (2015a,b), Ferreira (2016) e Gaspar (2015) descrevem sistemas de tradução de português para LGP, já referindo os autores ferramentas ligadas ao processamento de língua natural na geração de LGP. No entanto, estes trabalhos são provas de conceito que apenas cobrem um conjunto mínimo de fenómenos. Assim, cremos que o trabalho aqui proposto é o primeiro que, com o objetivo de desenvolver um tradutor para LGP (e não para português gestuado) tira verdadeiro partido de um corpus de LGP.

Em relação à representação dos gestos, não existe uma notação oficial para transcrever as componentes manuais e não manuais dos gestos. Existem diversos sistemas de escrita que variam entre simbólicos e textuais. Exemplos desses sistemas são o HamNoSys (sistema de notação de Hamburgo) (Hanke, 2004), sistema de notação de Stokoe (Stokoe, 2005), o Signwriting (Costa & Dimuro, 2003), o Sistema de Escritura Alfabética da língua gestual espanhola (SEA) (Herrero, 2003) e a glosa, que é a representação textual mais comum, sendo os gestos anotados usando as palavras com o mesmo significado na língua falada mas em letras maiúsculas (Mineiro & Colaço, 2010). Por exemplo, o gesto referente a coelho será representado pela glosa COELHO. Em 2015, foi estudada a aplicação do sistema Signwriting na LGP (Pinto, 2015).

4. Construção das regras de tradução e dicionário bilingue

Nas próximas subsecções apresentam-se os dados usados na construção das regras de tradução e descrevem-se as principais etapas que resultam na gramática de tradução usada pelo tradutor. Começamos por referir o corpus de referência (Secção 4.1); seguidamente descrevemos os pré-processamentos efectuados (Secção 4.2) antes de avançarmos para a tarefa de alinhamento (Secção 4.3). Finalmente, na Secção 4.4, descrevemos os recursos linguísticos obtidos, nomeadamente as regras de tradução e o dicionário bilingue, e na Secção 4.5 explicam-se as regras puramente manuais criadas para este trabalho.

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

Tabela 1: Convenções usadas na anotação de informações gramaticais no corpus.

Fenómenos Gramaticais	**Convenção (exemplo)**
Datilologia	`DT(M-A-R-I-A)`
Flexão em género	*`FG(MULHER+GATO)`*
Pron. Possessivos	*`PP(MEU)`*

Tabela 2: Exemplos de convenções usadas na anotação de fenómenos linguísticos no corpus de referência.

4.1. Corpus de referência

O corpus em desenvolvimento pela Universidade Católica Portuguesa é constituído por vídeos de surdos portugueses de diferentes faixas etárias (dos 10 aos 60 anos) e de diferentes regiões, 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 vídeo e áudio. Neste corpus, estão a ser anotados (entre outros):

- a tradução da mensagem enunciada no vídeo para português;
- os gestos (transcritos em glosas) e as respetivas classes gramaticais;
- os argumentos da frase: argumentos internos (complementos do predicado) e externos (sujeito da frase);
- o tipo de cada frase (interrogativa (`INT`), negativa (`NEG`) e exclamativa (`EXCL`) e por omissão, declarativa afirmativa).

Na anotação destas informações foram seguidas convenções. Na Tabela 1 descrevem-se algumas das convenções usadas na anotação de informações gramaticais e na Tabela 2 podem encontrar-se exemplos das convenções seguidas na anotação de alguns fenómenos linguísticos.

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

Atualmente, os dados utilizados na construção da gramática apresentada neste trabalho provêm de um vídeo de 5 minutos de um gestuante nativo com discurso informal e espontâneo. Das 66 frases que constituem estes dados, 3 são declarativas negativas, 5 interrogativas e as restantes declarativas afirmativas. A distribuição das principais classes gramaticais presentes nas frases em LGP encontra-se na Tabela 3. Na Tabela 4 apresentam-se estatísticas do vocabulário e das frases em português e em LGP do corpus. O corpus ainda está em desenvolvimento e novas regras podem ser geradas à medida que o corpus vai crescendo.

Classe gramatical	**Frequência**
Nomes	39.0%
Advérbios	7.3%
Verbos	22.0%
Adjetivos	8.5%
Numerais	5.6%
Pronomes	6.6%
Conjunções	4.4%

Tabela 3: Frequência das principais classes gramaticais nas frases analisadas.

	Português	**LGP**
Voc. total	575	412
Voc. único	280	221
Comprimento médio	8.7	6.2

Tabela 4: Estatísticas do vocabulário e do comprimento das frases em português e em LGP do corpus.

4.2. Fase de análise

Do corpus de referência apenas se conhecem as informações gramaticais das frases em LGP, pelo que as frases em português são analisadas sintática e morfossintaticamente através de ferramentas de processamento da língua natural. Num estudo preliminar determinaram-se as ferramentas que melhor levaram a cabo estas tarefas. Para a análise sintática, tratou-se do SpaCy (Honnibal & Montani, 2017); para a análise morfossintática, o FreeLing (Padró & Stanilovsky, 2012; Padró, 2012), sendo que este não providencia uma análise de dependências, mas uma análise morfossintática de maior granularidade e qualidade. Assim, 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 identificados através do FreeLing. Este último passo é realizado tanto nas palavras como nos gestos por ser a base do alinhamento de palavras e gestos descrito na Secção 4.3. Na análise sintática, a frase em português é dividida nos seus elementos frásicos (sujeito, predicado e modificador de frase), com base nas relações de dependência identificadas pelo SpaCy.

No final desta fase, as etiquetas resultantes da análise morfossintática são convertidas nas etiquetas do corpus de referência. 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 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`.

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 (Sousa, 2012). O seu tratamento foi deixado como trabalho futuro.

Por exemplo, dadas as frases em português, "*A Maria lê um livro*" e em LGP, MARIA LIVRO LER, no final desta fase conhecem-se as suas ordens frásicas e classes gramaticais, apresentadas em seguida.[5]

(1) A Maria lê um livro.

(2) MARIA LIVRO LER

(3) Ordem frásica da frase em português: SVO

(4) Ordem frásica da frase em LGP: SOV

(5) Classes gramaticais da frase em português: `N V N`

(6) Classes gramaticais da frase em LGP: `N N V`

A etiqueta `V` representa verbo, `N` corresponde a substantivo e `ADJ` é adjetivo, seguindo as convenções do corpus na Tabela 1.

Tendo as informações gramaticais das frases de ambas as línguas, reunem-se assim as condições para construir as regras de tradução.

[5] As informações gramaticais apresentadas são exemplificativas.

4.3. Alinhamento

Antes de passar à construção da gramática, há que alinhar o léxico das frases do corpus. Em sistemas de tradução estatísticos, o alinhamento do léxico é usualmente calculado através de métodos probabilísticos (Tambouratzis et al., 2012; Chiu et al., 2007; Sánchez-Cartagena et al., 2016), contudo no caso do par de línguas português-LGP, não existe um corpus suficientemente grande para treinar o alinhamento entre palavras e gestos. Assim, propomos um método baseado em medidas de semelhança (*string matching* e semelhança semântica), que se descreve de seguida.

É importante realçar que as correspondências entre uma palavra e um gesto não são simplesmente um-para-um. A Figura 2 esquematiza outros tipos de relações possíveis no alinhamento. A última relação é difícil de identificar, pelo que o seu tratamento foi excluído do âmbito deste trabalho.

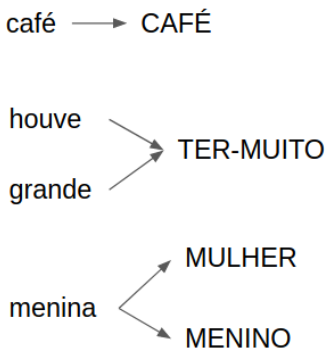

Figura 2: Os diferentes tipos de alinhamentos entre palavras e glosas. A primeira representa uma correspondência um-para-um, a segunda muitos-para-um e a última, um-para-muitos.

O alinhamento que propomos é o seguinte: as palavras e gestos são comparados letra-a-letra; se forem iguais são alinhados; caso contrário são comparados recorrendo à *OpenWordNet-PT* e, depois, a *word embeddings*. Esta última etapa vem reforçar o alinhamento semântico, pois se alguns pares palavra–gesto que não são alinhados pela WordNet, poderão sê-lo através de *word embeddings*.

A *OpenWordNet-PT*, por estar integrada na biblioteca NLTK e por oferecer várias medidas de semelhança entre dois conceitos,[6] foi a usada para calcular a semelhança semântica entre uma palavra e um gesto. Uma das medidas de semelhança é a semelhança de Wu-Palmer.[7] Considerou-se que uma palavra 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 medida de semelhança é apenas válida entre conceitos com a mesma classe gramatical, dado que não existe um hiperónimo comum entre *synsets* de diferentes classes gramaticais (Farkiya et al., 2015). Assim, adicionou-se outra premissa: uma palavra e um gesto são também semanticamente semelhantes se possuírem sinónimos com radicais semelhantes, como as palavras *arte* e *artístico*. 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.[8] Se para uma palavra e um gesto existir um par de sinónimos com valor dessa medida maior do que 0.8, então, essa palavra e esse gesto são alinhados. Caso contrário, passa-se para a etapa seguinte.

Quanto aos *word embeddings*, Hartmann et al. (2017) avaliam 31 modelos[9] de *word embeddings*[10] para português do Brasil e europeu. A avaliação revelou que para a analogia semântica e para português europeu, o modelo com melhor desempenho é o treinado com o algoritmo GloVe com 600 dimensões. Este modelo converte a palavra e o gesto em vetores. A semelhança entre as duas palavras relaciona-se com o ângulo formado pelos seus vetores, calculada através de a similaridade do cosseno:[11] quanto menor for o ângulo entre os vetores, maior é a semelhança entre as palavras. Se a palavra e o gesto tiverem um valor de semelhança maior do que 0.3, então são alinhados.

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. A Figura 3 exemplifica o resultado do alinhamento dos predicados das frases (1) e (2). Os determinantes dos predicados foram removidos na fase de análise.

[6] para alternativas, em (Oliveira et al., 2015) é feito um levantamento de bases de dados lexicais com relações semânticas entre palavras disponíveis para português e das suas características.

[7] Descrita em `https://www.nltk.org/howto/wordnet.html`.

[8] A biblioteca pyjarowinkler para Python foi usuada para o cálculo da distância de Jaro-Winkler.

[9] Encontram-se disponíveis em `http://nilc.icmc.usp.br/embeddings`.

[10] *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 (Hartmann et al., 2017).

[11] Detalhes sobre esta medida podem ser encontrados em `https://www.sciencedirect.com/topics/computer-science/cosine-similarity`.

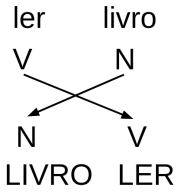

Figura 3: Resultado do alinhamento dos predicados das frases "*A Maria lê um livro.*" e MARIA LIVRO LER.

Os limites de semelhança usados nos passos anteriores foram decididos com base nos resultados das diferentes medidas de semelhança aplicadas a 36 863 pares palavras–gestos de outros vídeos do corpus, que contêm, apenas, transcrições em glosas.

O pseudo-código do alinhamento está descrito no Algoritmo 1. Neste processo de alinhamento usamos os lemas dos gestos e das palavras, o que permite alargar o número de correspondências exatas apanhadas pela primeira etapa. Por exemplo, as formas verbais *vão* e *ir* resultarão numa correspondência exata ao serem convertidos nos seus lemas (*ir* para ambos).

4.4. Regras de tradução e dicionário

Como se disse, do corpus resultam as regras de tradução e um dicionário bilingue.

Em sistemas de tradução estatística existem diferentes algoritmos para se extrair regras de tradução através do alinhamento de um corpus, como o algoritmo *phrase extraction*[12]. Neste tradutor adotou-se uma abordagem mais simples para construir as regras de tradução, explicada em seguida.

As regras de tradução dividem-se em dois tipos: as que descrevem a 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. As ordens frásica e dos constituintes morfossintáticos podem ser alteradas conforme o tipo de frase. Este fenómeno é comum 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 de tradução também são agrupadas de acordo com o tipo da frase (declarativa afirmativa, negativa, interrogativa e exclamativa) que originou a regra.

As regras de tradução descrevem as transformações gramaticais necessárias para que uma frase em português possa ser convertida na frase em LGP e, por isso, são compostas por dois "lados", nomeadamente pelo *lado português* e o *lado da LGP*. Os exemplos de regras dados daqui em diante seguem a estrutura descrita em (7):

(7) lado português → lado da LGP

As *regras frásicas* construíram-se a partir das ordens frásicas de cada frase em português, dadas pela análise sintática, e da ordem frásica da respetiva frase em LGP extraída do corpus. Por exemplo, considere que a frase em português em (1), "*A Maria lê um livro*" e a sua tradução em (2), MARIA LIVRO LER provêm do corpus. Da fase de análise (Secção 4.2) conhece-se a ordem frásica da frase em português (SVO) e do corpus sabe-se que a ordem frásica da frase em LGP é SOV. Com estas informações constrói-se diretamente a regra frásica em (8), respeitante a uma frase declarativa afirmativa.

(8) SVO → SOV

A construção das *regras morfossintáticas* baseia-se nas classes gramaticais dos elementos que compõem os pares palavra–gesto 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*, que permite identificar exatamente o que deve ser traduzido em quê. Por exemplo, do alinhamento na Figura 3, o par ler–LER contribui com um verbo para a regra morfossintática, ao qual se atribui o número de correspondência 1 (V1). Por sua vez, o par livro–LIVRO contribui com um nome (N2). Assim, forma-se a regra em (9). O mesmo se realiza para os restantes elementos frásicos. Os números de correspondência permitem preservar a troca de ordem entre o nome e o verbo.

(9) V1 N2 → N2 V1

Em (10) encontra-se mais um exemplo de uma regra morfossintática de um *predicado*, que determina a troca do constituinte *N2* para o fim da frase. De notar que sem os números de correspondência não seria possível determinar a correspondência entre os *N*s.

(10) V1 N2 ADJ3 N6 → V1 ADJ3 N6 N2

[12] O algoritmo *phrase extraction* é explicado em http://statmt.org/book/slides/05-phrase-based-models.pdf

Algoritmo 1: Alinhamento de palavras e gestos.

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

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.

Durante a construção das regras de tradução, procedeu-se à contagem da ocorrência de cada regra, para cada tipo de frase. Como se verá, estas estatísticas serão usadas no módulo de tradução (Secção 5). Além da sua importância no tradutor, apresentam informações linguísticas relevantes para o estudo de alguns fenómenos gramaticais da LGP, como a ordem canónica ou base.

Quanto ao dicionário bilingue de português e LGP, este foi construído automaticamente, com base no alinhamento das palavras com os gestos do corpus. Este recurso permite auxiliar a transferência lexical no tradutor (Secção 5.2), 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), existindo ainda pares semanticamente relacionados, como *religião* e IGREJA. Este dicionário foi posteriormente revisto com base nas informações transcritas do vídeo. Após a revisão e eliminação de correspondências erróneas como *século* e ARTE, o dicionário apresenta 102 entradas.

4.5. Regras manuais

Um conjunto de regras manuais complementa as regras de tradução anteriormente descritas. De notar que, num cenário em que as regras de tradução automática provêm de um corpus muito maior, estas regras seriam, provavelmente, desnecessárias.

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 as características da LGP. Integram também particularidades da língua relacionadas com a morfologia das palavras como a marcação do género feminino, dos tempos verbais e do grau do substantivo, assim como as expressões faciais gramaticais relativas às frases negativas e interrogativas.

Apesar de alguns fenómenos gramaticais da LGP estarem bem delineados, outros não o estão, como a marcação da negação. Existem várias formas de marcar a negação que variam tanto na expressão facial como no gesto manual, dependendo do verbo. Na pesquisa realizada para este artigo, não se encontraram estudos que indiquem em que contexto se recorre a cada uma das opções de marcação da negação. Deste modo, neste tradutor este fenómeno é tratado pela adição do marcador não manual *headshake* em simultâneo à componente manual NÃO, por ser o marcador manual mais frequente na LGP (Carmo et al., 2017), como está exemplificado em (11).

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

Para marcar as expressões faciais criou-se uma notação que identifica a expressão facial em si e a sua duração. A duração é identificada por chavetas: a chaveta aberta indicia o início 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, a frase (11) seria representada no tradutor como AMANHÃ DT(C-A-R-O-L-I-N-A) VESTIR {NÃO}(HEADSHAKE), o gesto não manual *headshake* é marcado por *(headshake)* e as chavetas indicam que este é produzido simultaneamente ao gesto manual de negação NÃO.

5. Tradutor

Nas próximas secções descrevem-se as fases da componente de tradução: primeiro, o pré-processamento (Secção 5.1), seguido das etapas de transferência lexical (Secção 5.2) e transferência sintática (Secção 5.3) e por fim a fase de geração morfológica (Secção 5.4). Os procedimentos de cada etapa serão exemplificados através da frase em (12) e dos seus elementos frásicos, *sujeito* em (13) e *predicado* em (14).

(12) A Diana perdeu o seu gatinho ontem.

(13) Sujeito: a Diana

(14) Predicado: perdeu o seu gatinho ontem.

5.1. Pré-processamento

A frase em português dada ao PE2LGP sofre um pré-processamento semelhante ao realizado no módulo de construção de regras de tradução (Secção 4): é analisada sintática e morfossintaticamente, os determinantes artigos (definidos e indefinidos), preposições e sinais de pontuação são removidos e as etiquetas resultantes das análises anteriores são convertidas para as do corpus, uniformizando-as com as das regras de tradução. 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 5.3).

5.2. Transferência lexical

O léxico português é mapeado no léxico da LGP com base no dicionário bilingue criado no módulo anterior. Caso a palavra esteja no dicionário, então será substituída pelo gesto correspondente; caso contrário, o seu lema será convertido em glosa na fase de geração (Secção 5.4). Admitindo que nenhuma das palavras da frase exemplo em (12) existe no dicionário bilingue, então esta não sofre alterações nesta fase.

5.3. Transferência sintática

A conversão da estrutura sintática da frase em português na correspondente estrutura sintática em LGP realiza-se pela aplicação das regras de tradução (Secção 4.4) e manuais (Secção 4.5). No caso das primeiras são aplicadas as que melhor se ajustam à estrutura sintática da frase em português conforme o tipo de frase. Para cada frase aplicam-se os dois tipos de regras de tradução, regras morfossintáticas e regras frásicas. É importante clarificar que as operações desta fase não se realizam sobre a frase em português mas sobre os seus elementos frásicos, divididos na análise sintática realizada no pré-processamento (Secção 5.1). Assim, o que é recebido nesta fase são as *estruturas sintáticas* de cada elemento frásico, exemplificadas em (15) para o sujeito e em (16) para o predicado da frase exemplo (os artigos definidos foram removidos no pré-processamento).

(15) Estrutura sintática do sujeito: N

(16) Estrutura sintática do predicado: V DET N ADV

A escolha da melhor *regra morfossintática* baseia-se no algoritmo da distância de edição (Wagner & Fischer, 1974) entre a estrutura sintática da frase de entrada e a estrutura sintática do lado português das regras morfossintáticas. A distância de edição é uma medida de semelhança entre duas sequências,[13] que permite saber que operações devem ser feitas para que as duas fiquem iguais. As operações possíveis são inserção, remoção e substituição. Os custos implementados para estas operações são de 1, exceto no caso em que o *tipo de frase* é substituído, ou seja, quando os tipos de frase da frase de entrada e da regra morfossintática são diferentes. Neste caso, o custo atribuído é 2, maior do que nas restantes operações, dado que a ordem dos constituintes morfossintáticos pode alterar-se consoante o tipo de frase. Desta forma diminui-se a probabilidade de a uma frase declarativa ser aplicada uma regra de uma frase interrogativa, por exemplo.

Antes de proceder-se ao cálculo das distâncias, tanto a estrutura da frase como a das regras do lado da língua portuguesa são convertidas para o formato em (17), 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`).

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

(17) CL1 CL2 CL3 Tipo_da_frase

Desta forma, as estrutura do sujeito e do predicado da frase exemplo são convertidas para:

(18) Sujeito: N CAN

(19) 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 do lado português e a frase. A regra a aplicar é a que apresenta menor distância entre a estrutura sintática da 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 no módulo anterior;
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 é consistente.

As regras de tradução que melhor se ajustam às estruturas sintáticas do sujeito e do predicado do exemplo estão indicadas respetivamente em (20) e (21). As distâncias obtidas foram de 0 para o sujeito e de 1 para o predicado.

(20) N1 CAN → N1 CAN

(21) V1 N2 ADV3 CAN → V1 ADV3 N2 CAN

A distância de edição, além da distância, indica as operações a realizar para tornar a estrutura sintática do lado português da regra igual à estrutura sintática da frase. As inserções no lado da LGP seguem uma heurística simples: o elemento a adicionar no lado da LGP é inserido a seguir à classe gramatical com o número de correspondência igual à classe gramatical anterior ao valor inserido no lado português. As operações de remoção e substituição 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. Por exemplo, para igualar as estruturas sintáticas do predicado em (19) e da regra em (21) basta inserir um *DET* depois do *V1* no lado português da regra e, seguindo a heurística anterior, no lado LGP deverá ser inserido um *DET* depois do constituinte morfossintático com o número de correspondência igual a 1, que é igualmente o constituinte *V1* e atribui-se ao novo constituinte o número de correspondência 4. Assim a transferência de estrutura sintática é determinada pela regra V1 DET4 N2 ADV3 → V1 DET4 ADV3 N2, que corresponde a *perdeu seu ontem gatinho*. A regra dita uma troca do constituinte *ADV3* (ontem) com o *N2* (gatinho).

Este procedimento garante que a todas as frases de entrada seja atribuída uma regra de tradução morfossintática.

De seguida, os elementos frásicos, com uma nova estrutura sintática, são unidos para formarem a frase em LGP. Esta união é baseada na ordem frásica mais frequente no corpus de acordo com o tipo da frase. Para frases declarativas afirmativas como a frase exemplo *A Diana perdeu o seu gatinho ontem.*, a ordem frásica mais frequente do corpus é SVO. Assim, os elementos frásicos são ordenados dessa forma, primeiro sujeito (*Diana*), depois verbo (*perdeu*) e no fim o objeto (*seu ontem gatinho*).

Contudo, e seguindo a premissa de estudos anteriores, em que se defende que a estrutura frásica base mais frequente da LGP é SOV, adicionou-se uma opção de escolha entre a estrutura mais frequente do corpus ou a estrutura SOV no tradutor. Se fosse escolhida esta estrutura, então o resultado da transferência sintática para a frase exemplo seria *Diana seu ontem gatinho perdeu*.

Por último, as regras manuais são aplicadas, através das quais os constituintes morfossintáticos são reordenados seguindo a gramática da língua. Dado que, em LGP, os advérbios de tempo são produzidos no início e os determinantes possessivos procedem o substantivo, o resultado desta fase da frase em (12) é *Ontem Diana perdeu gatinho seu*.

5.4. Fase de geração

Aqui, o léxico é convertido em glosas e são aplicadas as regras manuais relacionadas com a morfologia na LGP, como a marcação do graus diminutivo e aumentativo em substantivos (Secção 4.5). Desta fase sai uma sequência de glosas com marcadores adicionais que identificam expressões faciais e palavras soletradas seguindo as convenções de anotação do corpus de referência. Assim, o resultado da tradução da frase *A Diana perdeu o seu gatinho ontem.* 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", de acordo com a Tabela 2.

6. Avaliação

Para avaliar a qualidade da tradução do sistema proposto conduziram-se duas avaliações, uma automática, comparando a tradução do sistema com um corpus de teste, e outra manual com base na opinião de peritos.

6.1. Avaliação automática

As traduções produzidas por diferentes configurações do sistema PE2LGP foram avaliadas e comparadas com as do sistema *baseline* (Secção 6.1.3) com base nos dados do corpus de teste (Secção 6.1.1). Os objetivos desta avaliação são: averiguar se a abordagem seguida permite captar fenómenos linguísticos, produzindo LGP e não apenas português gestuado e perceber o impacto das regras de tradução na qualidade das traduções.

6.1.1. Corpus de teste

O corpus de teste foi criado por uma intérprete de português e LGP. É composto por 58 frases simples em português (em média com 5 palavras) e as correspondentes traduções em LGP, diferentes das do corpus de referência. O corpus será de domínio aberto. Para algumas frases em português foram anotadas mais do que uma tradução possível, mas não se procurou obter todas as traduções possíveis. Das 58 frases, 57 correspondem às formas negativas, interrogativas e declarativas afirmativas de 19 frases. A frase restante é a saudação "Bom dia".

Na Tabela 5 apresentam-se informações estatísticas sobre este corpus.

	Português	**LGP**
Voc. total	288	204
Voc. único	67	62
Comprimento médio	5.0	3.5

Tabela 5: Estatísticas do vocabulário e do comprimento das frases em português e em LGP do corpus.

6.1.2. Medidas de avaliação

As 58 frases em português do corpus de teste foram traduzidas pelo sistema e o seu resultado foi avaliado usando as medidas *Bilingual Evaluation Understudy* (BLEU) (Papineni et al., 2002) e *Translation Error Rate* (TER) (Snover et al., 2006). Os valores de BLEU apresentados são os valores cumulativos ao nível do corpus para 1-grama e 2-grama. Os valores variam entre 0 e 1, em que 1 assinala uma correspondência exata entre a *hipótese* (tradução do sistema) e a *referência* (tradução presente no corpus de teste). A medida TER corresponde à proporção de operações de edição a realizar para igualar a hipótese à referência, 0 indica uma correspondência exata. Os valores de TER apresentados são a média de TER de cada frase.

Para a medida TER foi ainda calculada a sua variância nas configurações dos conjuntos 1 e 2. Os valores de BLEU foram calculados sobre o corpus e não sobre frases individuais, por isso não apresentamos a sua variância.

6.1.3. Configurações

O sistema *baseline* consiste na produção de português gestuado. As frases traduzidas seguem a gramática do português e não possuem expressões faciais. Por exemplo, a tradução para português gestuado da frase *Quem comeu o bolo?* é QUEM COMER BOLO.

Distinguimos ainda o sistema aqui proposto de um sistema baseado puramente nas regras manuais. De notar que as frases em LGP que saem destes sistemas 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 anotado (SVO). Assim, no total conduziram-se 5 experiências, dispostas na Tabela 6. A configuração I é do sistema baseline, as configurações II e III pertencem ao sistema baseado apenas nas regras manuais e formam o *conjunto 1*, por fim, as configurações IV e V são do sistema proposto e formam o *conjunto 2*.

6.1.4. Resultados

A Tabela 7 apresenta os resultados para as medidas TER e BLEU das configurações dos vários sistemas. Os melhores resultados foram obtidos pelas traduções com a estrutura SOV traduzidas pelo sistema proposto e pelo sistema baseado somente em regras manuais (configurações II e IV). A Tabela 8 mostra a variância das diferentes configurações de cada conjunto.

6.1.5. Discussão dos resultados

Sistema baseline vs. restantes
Os resultados do sistema desenvolvido superaram os do sistema baseline, atingindo 0.29 de TER e 0.77 de BLEU para a estrutura SOV. Estes valores mostram que a aplicação das regras de

Configuração	Procedimento
Baseline	
I	SVO
Conjunto 1 – apenas regras manuais	
II	Estrutura SOV
III	Estrutura segundo o corpus de referência
Conjunto 2 – regras automáticas e manuais	
IV	Estrutura SOV
V	Estrutura segundo o corpus de referência

Tabela 6: Configurações experimentais.

Configuração	TER	BLEU	
		1-grama	2-gramas
Baseline			
I	0.86	0.5	0.13
Conjunto 1 – apenas regras manuais			
II	0.3	0.75	**0.64**
III	0.4	0.75	0.47
Conjunto 2 – regras automáticas e manuais			
IV	**0.29**	**0.77**	**0.64**
V	0.4	**0.77**	0.49

Tabela 7: Resultados das 5 configurações experimentais.

Conjuntos		Variância
1	II	0.10
	III	0.10
2	IV	0.09
	V	0.10

Tabela 8: Variância da medida TER das configurações dos conjuntos 1 e 2.

tradução e de regras manuais na transferência gramatical melhoram consideravelmente a qualidade das traduções, produzindo LGP e não português gestuado.

Conjunto 1 vs. conjunto 2

Os resultados de TER e BLEU entre as configurações que pertencem ao conjunto 1 e aquelas que pertencem ao conjunto 2 apresentam ligeiras diferenças. O mesmo se verifica quanto à variância entre os dois conjuntos. Esta proximidade entre os valores dos dois conjuntos deve-se à maioria das regras morfossintáticas aplicadas às frases declarativas afirmativas e negativas não alterarem a estrutura sintática da frase e pelo facto de as frases no corpus de teste possuírem estruturas sintáticas e morfossintáticas semelhantes, implicando a aplicação de regras semelhantes. Contudo, verificou-se que a aplicação das regras automáticas melhorou a qualidade de 2 traduções, igualando-as à referência. 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.

A comparação das traduções do sistema com as referências permitiu inferir que os erros nas traduções devem-se a: a) falhas na análise morfossintática; por exemplo, o verbo *quer* na frase *O segurança quer respeito?* foi classificado como uma conjunção coordenativa; b) limitações na identificação dos elementos frásicos e c) às regras morfossintáticas por descreverem apenas a

ordem das classes gramaticais principais. Esta última limitação implica que não sejam captados fenómenos relativos à ordem de determinados constituintes como os advérbios (ADV). Considerem-se os seguintes casos:

(22) li muito

(23) li ontem

Para os predicados em (22) e (23) a regra morfossintática a aplicar será a mesma por terem a mesma estrutura sintática (`V ADV`), admitindo que possuem o mesmo tipo de frase. Contudo, os dois advérbios são produzidos em ordens diferentes na LGP, *ontem* por ser um advérbio de tempo deverá ser produzido em primeiro, o que não acontece com o advérbio de quantidade *muito*. Este é um exemplo simples e ilustrativo, que seria possível resolver com regras manuais, mas se as regras morfossintáticas fossem mais finas conseguiriam tratar muitos destes casos só por si (que são numerosos e muitas vezes complexos para se resolver com regras manuais).

6.2. Avaliação manual

As métricas BLEU e TER usadas na avaliação automática podem não refletir a qualidade semântica da tradução produzida pelo sistema (Dorr et al., 2011; Snover et al., 2006) por serem medidas baseadas na correspondência exata entre o léxico das traduções e o das referências, sem considerar possíveis relações de sinonímia entre eles.

O objetivo desta avaliação é saber se o significado da frase em português prevalece na tradução, mesmo havendo diferenças na gramática e léxico em relação à referência. Assim escolheram-se 11 frases da avaliação automática que possuem diferenças significativas de léxico e de ordem das glosas que poderão afetar a compreensão da frase. A avaliação foi realizada com 4 peritos em linguística e com conhecimentos de LGP e português, a quem foram apresentadas sequências de glosas e pedido que as traduzissem para português (para avaliar se o significado da frase foi preservado na tradução do sistema) e que as classificassem quanto à qualidade da tradução das frases através de uma escala *Mean Opinion Score* (MOS) (Streijl et al., 2016), em *pobre*, *justo* e *bom*. *Pobre* quando 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.

As sequências de glosas apresentadas aos participantes correspondem a traduções produzidas pelo sistema PE2LGP segundo as regras manuais e as regras de tradução do corpus de referência (configuração V), por ser a configuração que usa todas as funcionalidades do sistema desenvolvido.

6.2.1. Resultados

A qualidade da tradução do presente sistema para 25% das frases foi *justa*, enquanto que para as restantes (75%) foi classificada como *boa*.

6.2.2. Discussão dos resultados

Os valores anteriores indicam que o significado da frase foi preservado em todas as traduções do sistema PE2LGP e 75% das traduções seguiram a gramática da LGP.

Os resultados da traduções de frases negativas destacam-se nesta avaliação por mostrarem problemas em todos os aspetos gramaticais (ordem frásica, ordem das glosas, expressões faciais e léxico). Em todas as frases negativas, os participantes indicaram que o verbo deveria ser colocado antes do gesto de negação ou simultâneo a ele, dependendo do verbo. Por exemplo, 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. Para 50% dos participantes o verbo TER foi considerado como um verbo copulativo, ou seja, deverá estar incorporado no objeto (OLHOS VERDES), ficando assim: NAMORADO MEU OLHOS VERDES {NÃO}(HEADSHAKE). Além das ordens dos constituintes este tipo de frases apresenta erros nos gestos manuais e nas expressões faciais. Contudo, não existe consenso sobre estes dois aspetos entre os participantes. Uns defendem que o gesto manual NÃO não é o indicado (mas sim o gesto NADA), outros afirmam que a negação é simultânea ao verbo e faz-se somente por expressão facial, e ainda que a expressão facial *headshake* não é a mais adequada para o dado contexto.

Nas 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, de novo, um consenso. 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): ocorre na última glosa (TER)

ou a partir da glosa PODER até ao final da frase.

Por fim, as observações feitas durante a entrevista pelos participantes indicam que a compreensão das sequências de glosas foi afetada pela ambiguidade lexical inerente às glosas e pela falta de contextualização das frases. Por exemplo, 3 dos 4 participantes interpretou a glosa SEGURANÇA em SEGURANÇA QUERER TAMBÉM RESPEITO como o sentimento de segurança e não a profissão de segurança. Este é um aspeto importante a ter em conta em avaliações de sequências de glosas.

7. Conclusões e trabalho futuro

A construção de um sistema de tradução de português europeu para LGP é condicionada pelos poucos recursos computacionais (e, no caso da LGP, linguísticos) disponíveis para estas línguas. A principal inovação deste tradutor face aos seus antecessores é a exploração do novo corpus em desenvolvimento pela Universidade Católica Portuguesa. Por norma, os tradutores desenvolvidos anteriormente utilizam exclusivamente regras de tradução manuais.

O novo corpus contém, além de anotações extensivas dos gestos utilizados e a sua tradução em português, informações gramaticais da LGP, como classes de palavras e expressões faciais e corporais. Assim, o sistema de tradução apresentado além de regras manuais, faz uso deste corpus anotado para gerar regras de tradução automática 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 e na transferência sintática, provocadas pela granularidade das regras morfossintáticas. De notar ainda que vários fenómenos associados à LGP não são ainda consensuais.

O estudo apresentado leva a crer que esta abordagem pode ser o ponto de partida para a criação de uma gramática computacional para a LGP, podendo o PE2LGP ser explorado em trabalho de investigação futuro, representando uma estratégia promissora no contexto atual dos recursos disponíveis para estas duas línguas. Uma avaliação com um corpus de teste com maior variabilidade de fenómenos gramaticais e com frases de diferentes complexidades deverá ser realizada para inferir o impacto das regras automáticas na qualidade das traduções. Vários fenómenos linguísticos ficaram ainda por tratar. Por exemplo, as preposições que requerem um tratamento apropriado.

Agradecimentos

Este trabalho foi parcialmente suportado pela Fundação para a Ciência e a Tecnologia através dos projectos UIDB/50021/2020 e PTDC/LLT-LIN/29887/2017, financiando este último a bolsa de Matilde Gonçalves.

Agradecemos a toda a equipa do projeto PTDC/LLT-LIN/29887/2017 da Universidade Católica Portuguesa, Helena Carmo, Mara Moita, Neide Gonçalves, Paulo Carvalho e Sebastião Palha, pela passagem de conhecimento sobre a língua gestual portuguesa. Um agradecimento especial a Neide Gonçalves pelo desenvolvimento do corpus de teste e da entrevista realizada na avaliação manual.

Referências

Almeida, Inês, Luísa Coheur & Sara Candeias. 2015a. Coupling natural language processing and animation synthesis in portuguese sign language translation. Em *Vision and Language 2015 (VL15), EMNLP 2015 workshop*, 94–103. 10.18653/v1/W15-2815.

Almeida, Inês, Luísa Coheur & Sara Candeias. 2015b. From European Portuguese to Portuguese Sign Language. Em *6th Workshop on Speech and Language Processing for Assistive Technologies (demo paper)*, 140–143. 10.18653/v1/W15-5124.

Amaral, Maria Augusta, Amandio Coutinho & Maria Raquel D. Martins. 1994. *Para uma gramática da língua gestual portuguesa* Colecção universitária. Caminho.

Araújo, Tiago, Felipe Ferreira, Danilo Silva, Leonardo Oliveira, Eduardo Falcão, Leonardo Domingues, Vandhuy Martins, Igor Portela, Yúrika Sato Nóbrega, Hozana Lima, Guido Lemos de Souza Filho, Tatiana Tavares & Alexandre Duarte. 2014. An approach to generate and embed sign language video tracks into multimedia contents. *Information Sciences* 281. 762–780. 10.1016/j.ins.2014.04.008.

Baltazar, Ana Bela. 2010. *Dicionário de Língua Gestual Portuguesa*. Porto Editora.

Bento, José. 2013. *Avatares em língua Gestual Portuguesa*. Lisbon, Portugal: Faculdade de

Ciências, Universidade de Lisboa. Tese de Mestrado.

Bettencourt, Maria Fernanda. 2015. *A ordem de palavras na língua gestual portuguesa: Breve estudo comparativo com o português e outras línguas gestuais*: Faculdade de Letras da Universidade do Porto. Tese de Mestrado.

Brour, Mourad & Abderrahim Benabbou. 2019. ATLASLang MTS 1: Arabic Text Language into Arabic Sign Language Machine Translation System. *Procedia computer science* 148. 236–245. 10.1016/j.procs.2019.01.066.

Bungeroth, Jan & Hermann Ney. 2004. Statistical sign language translation. *Workshop on Representation and Processing of Sign Languages, 4th International Conference on Language Resources and Evaluation, (LREC)* 105–108.

Carmo, Helena. 2016. *Uma primeira abordagem aos classificadores da língua gestual portuguesa*: Universidade Católica Portuguesa, Lisboa. Tese de Mestrado. http://hdl.handle.net/10400.14/22600.

Carmo, Helena, Verónica Milagres da Silva & Elsa Martins. 2017. Os verbos em negação na língua gestual portuguesa. *Cadernos de Saúde* 9. 15–25.

Chiu, Yu-Hsien, Chung-Hsien Wu, Hung-Yu Su & Chih-Jen Cheng. 2007. Joint optimization of word alignment and epenthesis generation for chinese to taiwanese sign synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(1). 28–39. 10.1109/TPAMI.2007.15.

Choupina, Celda. 2017. Aspetos estruturantes da morfossintaxe da LGP: expressão da quantidade e das categorias de sexo dos referentes animados. *Revista Leitura* 1. 4–25. 10.28998/2317-9945.2017v1n58p4-25.

Choupina, Celda, Ana Maria Barros Brito & Fernanda Bettencourt. 2016. Particularidades da morfossintaxe das construções ditransitivas com o verbo 'dar' na língua gestual portuguesa. *Revista da Associação Portuguesa de Linguística* 117–147. 10.21747/2183-9077/rapl2a6.

Choupina, Celda, Ana Maria Brito & Fernanda Bettencourt. 2017. Morphosyntax aspects of ditransitive constructions with the verb dar 'to give' in portuguese sign language. *Linguística : Revista de Estudos Linguísticos da Universidade do Porto* 11. 91–116.

Costa, Antonio da Rocha & Graçaliz Dimuro. 2003. SignWriting and SWML: Paving the way to sign language processing. Em *Atelier Traitement Automatique des Langues des Signes (TALN)*, s/p.

Davydov, Maksym & Olga Lozynska. 2017. Information system for translation into Ukrainian sign language on mobile devices. Em *12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, 48–51. 10.1109/STC-CSIT.2017.8098734.

Dorr, Bonnie, Joseph Olive, John McCary & Caitlin Christianson. 2011. Chapter 5: Machine translation evaluation and optimization. Em *Handbook of Natural Language Processing and Machine Translation*, 745–843. Springer.

Escudeiro, Paula, Nuno Escudeiro, Rosa Reis, Maciel Barbosa, José Bidarra, Ana Bela Baltasar, Pedro Rodrigues, Jorge Lopes & Marcelo Norberto. 2014. Virtual sign game learning sign language, 29–33.

Escudeiro, Paula, Nuno Escudeiro, Rosa Reis, Maciel Barbosa, José Bidarra, Ana Bela Baltazar & Bruno Gouveia. 2013. Virtual sign translator. Em *International Conference on Computer, Networks and Communication Engineering (ICCNCE)*, 290–292.

Escudeiro, Paula, Nuno Escudeiro, Rosa Reis, Jorge Lopes, Marcelo Norberto, Ana Bela Baltasar, Maciel Barbosa & José Bidarra. 2015. Virtual Sign–a real time bidirectional translator of Portuguese Sign Language. *Procedia Computer Science* 67. 252–262. 10.1016/j.procs.2015.09.269.

Farkiya, Alabhya, Prashant Saini, Shubham Sinha & Sharmishta Desai. 2015. Natural language processing using NLTK and WordNet. *International Journal of Computer Science and Information Technologies* 6.

Ferreira, António Vieira (ed.). 1997. *Gestuário: Língua Gestual Portuguesa*. SNR.

Ferreira, Rui. 2016. *PE2LGP 3.0: from european portuguese to portuguese sign language*: Instituto Superior Técnico, Universidade de Lisboa. Tese de Mestrado.

Forcada, Mikel L, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* 25(2). 127–144. 10.1007/s10590-011-9090-0.

Gameiro, João, Tiago Cardoso & Yves Rybarczyk. 2014. Kinect-Sign, teaching sign language to 'listeners' through a game. *Procedia Technology* 17. 384–391. 10.1016/j.protcy.2014.10.199.

Gaspar, Luís. 2015. *IF2LGP-Intérprete automático de fala em língua portuguesa para língua gestual portuguesa*: Instituto Politécnico de Leiria, Leiria. Tese de Mestrado.

Hanke, Thomas. 2004. HamNoSys-representing sign language data in language resources and language processing contexts. Em *Language Resources Evaluation Conference (LREC)*, s/p.

Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues & Sandra Aluisio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv:1708.06025* .

Herrero, Ángel. 2003. *Escritura alfabética de la lengua de signos española: once lecciones*. San Vicente del Raspeig: Publicaciones de la Universidad de Alicante.

Honnibal, Matthew & Ines Montani. 2017. SpaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Lima, Manuella et al. 2015. *Tradução automática com adequação sintático-semântica para LIBRAS*: Universidade Federal da Paraíba. Tese de Mestrado.

Luqman, Hamzah & Sabri A Mahmoud. 2018. Automatic translation of Arabic text-to-Arabic sign language. *Universal Access in the Information Society* 10.1007/s10209-018-0622-8.

Martins, Mariana & Ana Isabel Mata. 2017. Conexões interfrásicas manuais e não-manuais em LGP: Um estudo preliminar. *Linguística: Revista de Estudos Linguísticos da Universidade do Porto* 11. 119–138.

Mesquita, Isabel & Sandra Silva. 2009. *Guia prático de língua gestual portuguesa*. Editora Nova Educação.

Mineiro, Ana & Dora Colaço. 2010. *Introdução à fonética e fonologia na LGP e na língua Portuguesa*. Universidade Católica Editora.

Mohamed Amine, Cheragui. 2012. Theoretical overview of machine translation. *CEUR Workshop Proceedings* 867. 160–169.

Nascimento, Sandra & Margarita Correia. 2011. *Um olhar sobre a morfologia dos gestos*. Universidade Católica Editora.

Oliveira, Hugo Gonçalo, Valeria de Paiva, Cláudia Freitas, Alexandre Rademaker, Livy Real & Alberto Simões. 2015. As wordnets do português. *Oslo Studies in Language* 7(1).

Othman, Achraf & Mohamed Jemni. 2011. Statistical sign language machine translation: from English written text to American Sign Language Gloss. *International Journal of Computer Science Issues* 8(3). 65–73.

Padró, Lluís. 2012. Analizadores multilingües en freeling. *Linguamática* 3(2). 13–20.

Padró, Lluís & Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. Em *Language Resources and Evaluation Conference (LREC 2012)*, 2473–2479.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. Em *40th Annual Meeting of the Association for Computational Linguistics*, 311–318. 10.3115/1073083.1073135.

Pinto, Jorge Manuel Ferreira. 2015. *O SignWriting como um sistema de escrita apropriado às línguas gestuais: um contributo para o desenvolvimento de competências de escrita do aluno surdo?*: Universidade do Porto, Faculdade de Psicologia e Ciências da Educação. Tese de Doutoramento.

Porta, Jordi, Fernando López-Colino, Javier Tejedor & José Colás. 2014. A rule-based translation from written Spanish to Spanish Sign Language glosses. *Computer Speech & Language* 28(3). 788–811. 10.1016/j.csl.2013.10.003.

Rodrigues, Rute Ana Ferreira. 2018. *Compreensão da língua gestual portuguesa em crianças surdas. proposta de um instrumento de avaliação*: Escola Superior de Educação Paula Frassinetti. Tese de Doutoramento.

San-Segundo, Rubén, Roberto Barra-Chicote et al. 2006. A Spanish speech to sign language translation system for assisting deaf-mute people. Em *9th International Conference on Spoken Language Processing, INTERSPEECH 2006 – ICSLP*, 17–21.

Sánchez-Cartagena, Víctor M, Juan Antonio Pérez-Ortiz & Felipe Sánchez-Martínez. 2016. RuLearn: an Open-source Toolkit for the Automatic Inference of Shallow-transfer Rules for Machine Translation. *The Prague Bulletin of Mathematical Linguistics* 106(1). 193–204.

dos Santos, Ruben. 2016. *PE2LGP: do texto à língua gestual*: Instituto Superior Técnico, Universidade de Lisboa. Tese de Mestrado.

Shieber, Stuart M & Yves Schabes. 1990. Synchronous tree-adjoining grammars. Em *13th Conference on Computational linguistics*, 253–258.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. Em *7th Conference of the Association for Machine Translation of the Americas*, 223—-231.

Sousa, Ana Paula de Almeida. 2012. *Interpretação da língua gestual portuguesa*: Faculdade de Ciências. Tese de Doutoramento.

Stokoe, William C., Jr. 2005. Sign language structure: An outline of the visual communication systems of the american deaf. *The Journal of Deaf Studies and Deaf Education* 10. 3–37. 10.1093/deafed/eni001.

Streijl, Robert C, Stefan Winkler & David S Hands. 2016. Mean Opinion Score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* 22(2). 213–227. 10.1007/s00530-014-0446-1.

Su, Hung-Yu & hung-Hsien Wu. 2009. 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* 17(7). 1305–1315. 10.1109/TASL.2009.2016234.

Tambouratzis, George, Michalis Troullinos, Sokratis Sofianopoulos & Marina Vassiliou. 2012. Accurate phrase alignment in a bilingual corpus for ebmt systems. Em *5th Building and Using Comparable Corpora Workshop*, vol. 26, 104–111.

Wagner, Robert A & Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)* 21(1). 168–173. 10.1145/321796.321811.

Zhao, Liwei, Karin Kipper, William Schuler, Christian Vogler, Norman Badler & Martha Palmer. 2000. A machine translation system from English to American Sign Language. Em John S. White (ed.), *Envisioning Machine Translation in the Information Future*, 54–67. Springer Berlin Heidelberg.

Propuesta recibida en octubre 2020 y aceptada para publicación en mayo 2021.

Corpus Paralelo de Español, Inglés y Chino y Análisis contrastivo del tiempo pasado del español a partir de corpus

Parallel corpus of Spanish, English and Chinese and corpus-based contrastive analysis of the past tense in Spanish

Hui-Chuan Lu
National Cheng Kung University

An Chung Cheng
University of Toledo

Meng-Hsin Yeh
National Cheng Kung University

Chao-Yi Lu
National Cheng Kung University

Ruth Alegre Di Lascio
National Cheng Kung University

Resumen

El presente estudio se dedica al desarrollo de un corpus paralelo trilingüe denominado CPEIC (Corpus Paralelo de Español, Inglés y Chino) cuyo fin es el de aportar conocimientos a las investigaciones sobre la traducción, el análisis contrastivo, el aprendizaje y la enseñanza de una lengua extranjera. Dicho CPEIC abarca las tres lenguas más habladas del mundo (español, inglés y chino) y contiene aproximadamente 4 millones de palabras. Basándose en el corpus paralelo desarrollado, se realizó un análisis contrastivo del tiempo pasado, el cual se expresa de manera diferente en las tres lenguas mencionadas. Los resultados obtenidos (a) avalan estudios previos sobre la relación entre el pretérito del español con el marcador aspectual chino "le", así como también la relación entre el imperfecto del español con "would" y "was/were+Ving" del inglés, (b) contradicen las presunciones con respecto a la conexión entre el imperfecto del español y el marcador aspectual chino "zhe", y (c) proporcionan una nueva perspectiva sobre la relación entre el pretérito del español y la voz pasiva en los tres idiomas.

Palabras clave

corpus paralelo, traducción, análisis contrastivo, tiempo pasado, aspecto, estudio a partir de corpus

Abstract

This study constructed a trilingual parallel corpus, the Parallel Corpus of Spanish, English and Chinese (CPEIC in Spanish), and used it to benefit research in translation, contrastive analysis, language learning, and teaching. The CPEIC contains the world's 3 most-spoken languages and comprises approximately 4 million words. Based on the construction result, a parallel corpus-based contrastive analysis was dedicated to the study of the past tense, which functions differently in the 3 languages. The results (a) support previous studies in associating the Spanish preterit and the Chinese aspectual marker "le", and in relating the Spanish imperfect with the English "would" and "was/were+Ving," (b) contradict assumptions for connecting the Spanish imperfect with the Chinese aspectual marker "zhe", and (c) offer new insights in uniting the Spanish preterit with the passive voice in all 3 languages.

Keywords

parallel corpus, translation, contrastive analysis, past tense, aspect, corpus-based

1. Introducción

La construcción del corpus y su respectiva investigación han jugado un papel crucial en el desarrollo de la lingüística durante los últimos 20 años. De acuerdo con los corpus enumerados en *Bookmarks for Corpus-based Linguistics* (Lee, 2010), la mayoría son corpus en inglés y fueron construidos con un propósito específico. Los corpus paralelos relacionados con lenguas distintas al inglés han atraído escasa atención. Esto puede ser atribuido a las dificultades de obtención y manejo de datos paralelos provenientes de varios idiomas. Sin embargo, debido a que la creación de un corpus paralelo involucra múltiples lenguas, puede proporcionar diversos valores académicos y de aplicación que no se encuentran en otros tipos de corpus. En comparación con los corpus monolingües o comparables, un corpus paralelo multilingüe puede ayudar al estudio de la traducción y al análisis contrastivo, con el fin de controlar las variables ocultas en el contenido no paralelo, y así llegar a una conclusión más objetiva y convincente.

DOI: 10.21814/lm.13.1.340

A diferencia del uso de las concordancias paralelas ya existentes o las herramientas del corpus, un corpus paralelo con funciones de búsqueda puede obtener resultados avanzados de manera más efectiva, como, por ejemplo, datos etiquetados y alineamiento de palabras, lo cual reduciendo así el tiempo de búsqueda. Entre los corpus paralelos existentes (Lee, 2010), no existe ningún corpus paralelo simultáneo entre los tres idiomas más hablados del mundo: inglés, español y chino, aunque cabe destacar que existen alineados de dos en dos (inglés–español, inglés–chino). Por lo tanto, la creación de un corpus paralelo trilingüe de este tipo facilitaría la investigación sobre la lingüística contrastiva y podría ser aplicada al estudio de la adquisición de una segunda lengua o el multilingüismo.

Basándose en el corpus paralelo trilingüe creado en este estudio, se examinó una característica lingüística en particular: el tiempo pasado. Este tiempo verbal se emplea de manera diferente en estos tres idiomas. El tiempo pasado en el español posee dos posibles conjugaciones para el verbo: el pretérito y el imperfecto, mientras que en el inglés sólo se cuenta con una conjugación para el verbo: el tiempo pasado ("past tense"). El chino, por su parte, no posee inflexiones verbales. Esta característica lingüística fue investigada debido a que el tiempo pasado en el español resulta ser una de las reglas gramaticales más difíciles de aprender para los estudiantes taiwaneses, cuya lengua materna (L1) es el chino, y cuya segunda lengua (L2) y primera lengua extranjera es el inglés. Por lo tanto, el corpus paralelo trilingüe construido refleja el contexto bajo el cual el idioma español es asimilado en Taiwán. La creación del corpus paralelo trilingüe, los hallazgos del estudio de la traducción y el análisis contrastivo utilizando el corpus construido, proporcionan pautas útiles para la traducción del español y la enseñanza o el aprendizaje de otros idiomas.

Este artículo está organizado de la siguiente manera: en la segunda sección, se revisan estudios previos y concurrentes relacionados al tema. En la tercera sección, se presenta la creación y evaluación del corpus paralelo trilingüe. En la cuarta sección, se provee un estudio de traducción y análisis contrastivo basado en el corpus paralelo construido y en sus implicaciones pedagógicas. Finalmente, en la quinta sección se brinda una conclusión del artículo.

2. Revisión literaria

2.1. Corpus Paralelo

Utilizando el término "corpus paralelo" como filtro para la búsqueda en la base de datos de *Linguistics and Language Behavior Abstracts* (LLBA) se observó el aumento de publicaciones de investigaciones relacionadas a dicho tema en los últimos 40 años. La gran mayoría de estos estudios se han enfocado en la construcción de un corpus paralelo. Por lo general, estos estudios se han centrado en lenguas indoeuropeas. En consecuencia, las combinaciones de lenguas asiáticas son relativamente escasas, encontrándose sólo la del japonés–chino (Ma et al., 2004).

En el proceso de creación del corpus paralelo, la alineación entre los diferentes idiomas resultó ser el problema más difícil de superar, igual que lo previamente mencionado por Sun et al. (2002). En la actualidad, la alineación de oraciones en el proceso de creación de un corpus paralelo no presenta el mismo nivel de complejidad que la alineación de palabras, cuyo proceso require de más pasos, pero que, a su vez, obtiene mejores resultados de correspondencia semántica. Con el fin de obtener un conocimiento general sobre el desarrollo actual de los corpus paralelos, se evaluaron aquellos relacionados a los tres idiomas más hablados del mundo: español, inglés y chino. Una vez evaluados los resultados, dos tipos de corpus paralelos fueron identificados. El primer tipo compila datos en paralelo en una carpeta, permitiendo la descarga de dichos datos; no obstante, nos enfocamos en el segundo tipo, el cual proporciona funciones de búsqueda y otras funciones más avanzadas.

La primera categoría incluye (1a) the European Parliament Proceedings Parallel Corpus, o, Corpus Paralelo de las Actas del Parlamento Europeo (Koehn, 2005) y (1b) Multilingual and Parallel Corpora European Commission, o, la Comisión Europea de Corpus Paralelos y Multilingües (MLCC por sus siglas en inglés) (ELRA, 1996). El primer corpus, (1a), contiene fuentes procedentes de las actas del Parlamento Europeo (1996–2009), incluyendo corpus paralelos del inglés (49.093.806 palabras) y español (51.575.748 palabras), con 1.965.734 oraciones paralelas en total. El propósito para la creación de dicho corpus fue el de desarrollar un sistema de traducción automática, el cual no dispone de ninguna función de búsqueda. Los datos del segundo corpus, (1b), se obtuvieron del Diario Oficial de las Comunidades Europeas (6.000.000–9.000.000 palabras para cada idioma), que cobra una tarifa por su uso. Los siguientes corpus cuentan

con interfaces de usuario y funciones de búsqueda de palabras: (2a) Corpus Paralelo de Análisis Contrastivo y Traducción Inglés–Español, P-ACTRES (Rabadán, 2002), que consta de libros, editoriales de periódicos, artículos de revistas y ensayos (2.500.000 palabras). El propósito para la creación de este corpus fue el de realizar un análisis contrastivo del inglés y el español, las aplicaciones del inglés como lengua extranjera y la enseñanza de la traducción. Las búsquedas de palabras y POS están disponibles. A diferencia de otros corpus, (2b) Open Source Parallel Corpus, OPUS (Tiedemann, 2012) ofrece una búsqueda pública de datos de traducción entre el español (400.000–500.000 palabras) y el inglés (500.000 palabras). Como resultado de dicha búsqueda se presenta la alineación de las oraciones provenientes de estos dos idiomas. Por su parte, (2c) English–Chinese Parallel Concordance, E–C Conco d (Lixun, 2001) contiene 1.878.795 palabras en inglés y 3.152.866 caracteres chinos, derivados de novelas, documentos legales, artículos académicos, cuentos de hadas, discursos, ensayos, y fábulas. El mismo dispone de funciones de búsqueda de palabras y como resultado se presenta la alineación de oraciones entre estos dos idiomas. (2d) El Corpus Paralelo Inglés–Chino de Babel, ParaConc (McEnery & Xiao, 2005) comprende 253.633 palabras en inglés y 287.462 caracteres chinos, y cuyas fuentes de datos son *World of English* y *Time*. Utilizando ParaConc, los resultados de búsqueda generados son oraciones alineadas etiquetadas con POS.

Algunos corpus paralelos ya existentes se encuentran en etapa de recopilación de información; sin embargo, ninguno trabaja simultáneamente con los tres idiomas, tampoco presentan una interfaz de usuario, ni consideran las búsquedas de palabras o las de POS. Esta notable falta de un corpus paralelo español–inglés–chino motivó la realización del presente estudio. Con base en la revisión de estudios previos (Corpas Pastor, 2003; Castillo Rodríguez, 2009; Baker, 1995; Malmkjær, 2005; Rabadán et al., 2009; Dimitrova et al., 2010), se construyó un corpus paralelo con el objetivo de realizar un análisis de contraste enfocado en las características lingüísticas del tiempo pasado del español, lo cual beneficiará a investigaciones futuras en el área del aprendizaje de idiomas.

2.2. Tiempo Pasado

Mediante un análisis de contraste basado en los corpus paralelos, se investigaron las conjugaciones de los tiempos pasados del español y sus interacciones y diferencias con las del inglés y del chino. Dichas diferencias se pueden observar en las siguientes tres oraciones:

(1) Cuando conversábamos, sonó el teléfono.
(2) When we were talking, the telephone rang.
(3) 當 我們正在聊天時， 電話 響了。
Dang women zhengzai liaotianshi dianhua xiangle

En el español, el tiempo pasado está marcado rotundamente; la morfología flexiva indica tanto el tiempo como el aspecto del verbo. El pretérito codifica la perfectividad, mientras que el imperfecto codifica la imperfectividad. Como se muestra en la oración (1), la terminación flexiva "-ábamos" es utilizada como marcador del imperfecto, y "-ó" es un marcador que simboliza el pretérito. En el inglés, las distinciones aspectuales de los verbos son menos notorias que las distinciones temporales utilizadas por los hablantes nativos del inglés. En el inglés, el contraste aspectual más evidente se encuentra entre el progresivo y el perfectivo al momento de utilizar el pasado progresivo y el pasado simple. En la oración (2), la terminación flexiva "-ing" indica la acción progresiva y "rang" indica el pasado simple. Por otro lado, en el chino no se marcan contrastes en tiempos y aspectos verbales a través de la morfología. Los hablantes nativos del chino son capaces de identificar y comprender los distintos tiempos verbales a través de adverbios, adjuntos, argumentos y referencias contextuales. En la oración (3), "zai" es un marcador aspectual que denota un significado progresivo y "le" es un marcador de aspecto perfectivo.

Con base en la literatura previa y la observación gramatical, las variables que juegan un papel en la interacción de los tiempos verbales entre estos tres idiomas incluyen: los aspectos gramaticales del español (el aspecto del pretérito y el marcador imperfecto), los objetos de los verbos, la negación, los adverbios y conjunciones temporales, los tiempos verbales del inglés (el pasado simple y progresivo) y los marcadores aspectuales del chino ("guo, zai, zhe, le") (Bardovi-Harlig, 2000; Robison, 1990; Salaberry, 2002, 2011; Vendler, 1967; Xiao & McEnery, 2004). Estas variables fueron examinadas con un énfasis especial en el corpus.

3. Metodología

3.1. Creación del CPEIC

Para diferenciarse y resaltar entre los corpus paralelos ya existentes, se establecieron los siguientes objetivos relacionados al idioma, la alineación, el etiquetado, la función de búsqueda y la presentación de resultados: (a) el corpus paralelo trilingüe de español–inglés–chino, (b) el etiquetado POS y la alineación de palabras y (c) la búsqueda simultánea en varios idiomas y a través de palabras clave. El procedimiento de creación consta de cuatro pasos principales: (a) la compilación de datos, (b) el etiquetado POS, (c) la alineación de palabras y (d) la programación de software, como se muestra en la Figura 1.

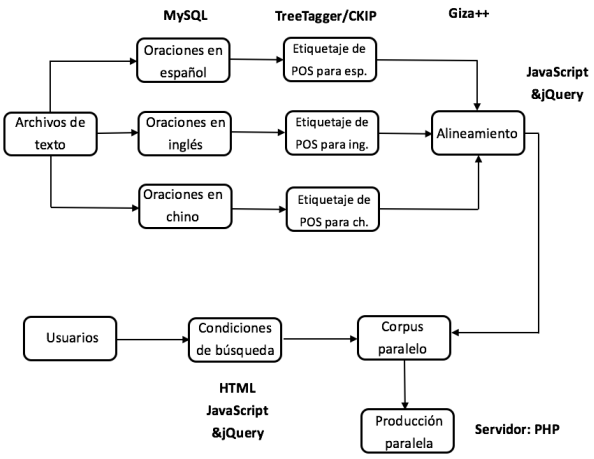

Figura 1: Diagrama de flujo de creación de CPEIC.

Se comenzaron a recopilar los datos paralelos para la CPEIC en el 2007 a partir de tres fuentes: (a) la Biblia, (b) cuentos de hadas y (c) formas escritas y habladas de documentos de las Naciones Unidas (ONU) en los tres idiomas (español, inglés y chino) con el objetivo de incluir un amplio repertorio de temas y de tipos de texto. Las fuentes encontradas fueron escasas debido a que existen pocos textos paralelos en español y chino que sean gratuitos y de fácil disposición en Internet. El manejo de datos varió según la fuente debido a que los textos que provienen de distintos sitios web cuentan con diversos formatos de sistema en Internet. Hasta la fecha, el CPEIC contiene 1.193.418 palabras en español, 1.199.715 palabras en inglés y 1.200.914 caracteres chinos.

En las etapas iniciales, los datos recopilados para el español e inglés se importaron a MySQL para etiquetarlos en POS utilizando TreeTagger (Solorio & Liu, 2008), mientras tanto, para el chino se utilizó el Procesamiento de Conocimiento e Información en Chino (CKIP), sistema de segmentación de palabras en chino (Ma & Chen, 2003). Posteriormente, las palabras se alinearon del idioma A al B, y seguidamente del idioma B al A utilizando Giza++ (Och & Ney, 2000). Este paquete de software procesa abundantes datos con el fin de minimizar los errores al alinear la lengua fuente y la lengua meta; y se basa en el modelo de probabilidad optimizado para predecir las correspondencias semánticas más posibles. Se diseñó una interfaz de usuario de entorno web utilizando JavaScript y jQuery, mientras que el servidor fue programado con el Preprocesador de Hipertexto (PHP). La interfaz de usuario se muestra en la Figura 2.

Figura 2: Interfaz de usuario CPEIC.

3.2. Evaluación del CPEIC

Tras culminar la primera etapa de la construcción del CPEIC, los resultados obtenidos fueron evaluados, como se indica en esta sección. De acuerdo con la evaluación realizada, el uso del corpus paralelo construido para la búsqueda de datos contrastivos puede ser siete veces más rápido que el tradicional. Al examinar los objetivos establecidos y comparar el corpus con los ya existentes, el CPEIC creado se destaca por las siguientes razones: (a) es un corpus paralelo que resuelve el problema de compatibilidad de tres idiomas, (b) los datos están etiquetados con POS y alineados con palabras para la función de búsqueda y presentación de resultados, y (c) la búsqueda en la interfaz de usuario puede ser simultáneamente trilingüe e incluir múltiples palabras clave y consultas compuestas con una velocidad de búsqueda mejorada y una frecuencia de cierre reducida.

Además, el corpus construido fue evaluado calculando el porcentaje de precisión del etiquetado POS de cada idioma y la alineación de palabras entre dos idiomas. Los resultados indicaron que las tasas de precisión del etiquetado POS para la Biblia, los cuentos de hadas y los documentos de la ONU escritos y orales excedieron el 93 % para inglés y español, mientras que variaron del 76 % al 93 % para el chino en las diversas

fuentes de datos. Esta diferencia entre la tasa de precisión de etiquetado del español e inglés con la del chino podría ser el resultado de los diferentes sistemas de etiquetado utilizados para las fuentes de datos: TreeTagger para inglés y español, y CKIP para chino. Si bien es cierto que este corpus paralelo trilingüe puede simplificar el proceso de búsqueda en una investigación proporcionando abundantes datos, múltiples funciones y un motor de búsqueda rápido, los sistemas adoptados mostraron errores en los resultados de búsqueda. En la etapa actual, los resultados parcialmente incorrectos en el etiquetado POS y la alineación de palabras han requerido una verificación posterior y corrección manual a fin de obtener mejores resultados.

Finalmente, evaluamos el valor del autoaprendizaje desde la perspectiva de los usuarios, identificando las ventajas de la aplicación del corpus construido. Entre estas ventajas se encuentra el poder diferenciar dos elementos similares al proporcionar niveles de oración y párrafo para que así los alumnos obtengan más detalles y entendimiento léxico. En cuanto a la satisfacción del usuario, según la experiencia de búsqueda, más del 69 % de los participantes informaron que la CPEIC fue útil para ayudarlos a adquirir conocimientos lingüísticos, en particular para proporcionar traducción paralela a través de similitudes o diferencias de estructuras y expresiones léxicas.

4. Estudio contrastivo del tiempo pasado basado en el corpus paralelo

4.1. Objetivo y preguntas de investigación

El estudio de la traducción y el análisis contrastivo mejoran nuestro conocimiento de la gramática universal y los parámetros específicos para cada idioma, lo cual también aumenta nuestra comprensión de la transferencia de estos. Debido a que L1 (lengua materna), L2 (primera lengua extranjera) y L3 (segunda lengua extranjera) son tres idiomas paralelos para los estudiantes taiwaneses de español, el estudio de la traducción y el análisis contrastivo de estos idiomas ayuda a examinar la influencia de la primera y segunda lengua en la adquisición de la tercera lengua. Por lo tanto, este estudio abordó la siguiente pregunta de investigación: Basado en el estudio de la traducción y el análisis contrastivo de un corpus paralelo trilingüe, ¿cuáles son las variables lingüísticas que podrían ser asociadas con los dos aspectos gramaticales del tiempo pasado en español?

4.2. Metodología

Al analizar los elementos léxicos y las estructuras sintácticas de los resultados alineados, las similitudes y diferencias se compararon y contrastaron observando los tres lenguajes paralelos y la interacción entre ellos. Con el objetivo de determinar los factores relevantes que podrían afectar la selección de dos aspectos gramaticales diferentes en el tiempo pasado del español, las posibles variables incluidas en el examen fueron las siguientes: los adverbios y las conjunciones temporales en español, las negaciones, los objetos de los verbos y la voz pasiva; el tiempo pasado en inglés (tiempo pasado simple, progresivo, y futuro en tiempo pasado) y voz pasiva; y cuatro marcadores de aspecto en chino ("guo, zai, zhe y le").

Los datos analizados se extrajeron del corpus trilingüe paralelo construido, (CPEIC), que incluía las tres fuentes antes mencionadas: la Biblia, los cuentos de hadas y las formas escritas y orales de los documentos de la ONU. En este estudio se examinaron un total de 198.386 palabras y se analizaron 2.160 verbos en tiempo pasado en español y los elementos correspondientes en inglés y chino.

4.3. Resultados y debate

Mediante la aplicación del análisis contrastivo de los datos paralelos provenientes de diversas fuentes, se obtuvieron los siguientes resultados. El resultado más relevante demostró la contradicción entre los marcadores de aspecto de acción progresiva en chino "zhe" con el imperfecto del español. El marcador de aspecto "zhe" se empleó de manera diferente a lo que se esperaba. La mayor correspondencia al marcador progresivo "zhe" del chino fue alcanzada por el pretérito del español y no por el imperfecto, en el caso de la biblia el pretérito se utilizó en un 80 % de los casos y en los cuentos de hada en un 59 %.

Por lo tanto, seguir la suposición convencional de que el marcador progresivo chino "zhe" está asociado con la forma imperfecta del español crearía una correspondencia errónea entre el español y el chino. Sin un estudio paralelo basado en un corpus, esta evidencia contraria no se habría descubierto. Por lo tanto, los hallazgos del análisis contrastivo translingüístico verifican el conocimiento gramatical previo.

Al continuar con el estudio paralelo se realizó otro descubrimiento destacado que estudios previos no han discutido. Entre los factores que inciden en la selección entre los dos aspectos del tiempo pasado del español, la voz pasiva en los tres idiomas (*ser*+PP(100 %) en español,

be+PP(100 %) en inglés, y *bei* (más del 67 %) en el chino) está altamente asociada al uso del pretérito en el español y no al imperfecto.

Además, el resultado proporciona evidencia basada en el corpus para respaldar el conocimiento gramatical previamente establecido. Entre las variables examinadas en inglés, el tiempo progresivo, "*be* + Ving" (más del 93 %) y el futuro en tiempo pasado "*would* + V" (más del 92 %) tienen una fuerte relación con el tiempo imperfecto español. En cuanto al chino, el marcador de aspecto perfectivo "le" tiene una alta tendencia (más del 82 %) a aparecer con el pretérito español.

También derivamos ciertos principios concluyentes para la enseñanza y el aprendizaje del español. En primer lugar, el pretérito en español (87 %, 60 %, 89 % y 88 % para la Biblia, los cuentos de hadas y las formas escritas y orales de los documentos de la ONU, respectivamente) generalmente aparece con más frecuencia que la forma imperfecta, lo que sugiere una mayor frecuencia del pretérito en el input del aprendizaje de este idioma. Además, Lu et al. (2015) propusieron que los estudiantes de segunda lengua (estudiantes de habla inglesa y mandarín) adquieran el pretérito español antes que el imperfecto. Por lo tanto, proponemos enseñar el tiempo pasado en español en función del orden de frecuencia de cada aspecto y la adquisición del alumno de dicho aspecto. Los estudiantes deben aprender los verbos de alta frecuencia de acuerdo con el siguiente orden de aspecto gramatical: (a) verbos utilizados exclusivamente en pretérito, (b) verbos utilizados exclusivamente en el imperfecto, (c) verbos utilizados con más frecuencia en pretérito que en el imperfecto, y (d) verbos utilizados con más frecuencia en el imperfecto que el pretérito. Además, ciertas palabras clave están relacionadas con aspectos específicos, como por ejemplo, la aparición de la conjunción temporal "mientras" y el adverbio temporal "siempre" garantizan el uso del aspecto imperfecto (100 %) en los textos de cuentos de hadas, mientras que la mayoría de las conjunciones temporales españolas, "cuando" o "when" (en inglés), se asocian con la forma pretérita en la Biblia (88 %), los cuentos de hadas (76 %) y en la forma oral de los documentos de la ONU (91 %).

Finalmente, además de las generalizaciones concluyentes antes mencionadas, la relación entre los factores y las fuentes del texto a veces presentan cambios. Por ejemplo, las negaciones *no* ("no") y *ni* ("neither") tienden a aparecer más frecuentemente con el imperfecto español en los cuentos de hadas y en los escritos de los documentos de la ONU, mientras que muestran comportamientos diferentes en la Biblia y en los documentos orales de la ONU. Por lo tanto, la selección de materiales didácticos para propósitos específicos debe tener en cuenta géneros y formas.

4.4. Enseñanza y aprendizaje

Un corpus paralelo trilingüe puede ser utilizado en investigaciones lingüísticas y presentar numerosas implicaciones pedagógicas para la traducción, enseñanza y aprendizaje de una segunda lengua para los estudiantes de estas lenguas. Con base en la traducción y los resultados contrastivos a través de búsquedas utilizando el corpus trilingüe construido, los profesores pueden diseñar materiales didácticos y llamar la atención de los estudiantes de español L3 para fortalecer los efectos positivos de L1 (chino) y L2 (inglés) y evitar sus efectos negativos según las similitudes y las diferencias entre las tres lenguas.

Para los principiantes, en comparación con los corpus monolingües, el presente corpus multilingüe se puede utilizar para buscar ejemplos o correlaciones que correspondan con la traducción de la lengua materna o la primera lengua extranjera para facilitar la determinación de significados (por ejemplo, *ser* o "be" vs. *estar* o "be"). Para estudiantes más avanzados, el corpus construido puede ayudar en la aclaración de las diferencias triviales entre las formas relacionadas mediante la búsqueda transversal de dos aspectos (por ejemplo, *fueron* vs. *eran* o *estaban* vs. *estuvieron*). Además, dado que, el CPEIC puede obtener de manera sistemática y eficiente patrones, características y categorías clasificadas basándose en la frecuencia del uso del pretérito y del imperfecto, sugerimos que los profesores consideren la frecuencia de aparición de los verbos en los datos del idioma nativo para diseñar sus materiales de enseñanza con verbos de alta frecuencia (*decir, salir, estar* "say, leave, be") siguiendo la secuencia de aprendizaje (sólo verbos en pretérito > sólo verbos en imperfecto > más pretérito que imperfecto > más imperfecto que pretérito) y obteniendo una lista de palabras (*soler* "usually do" en la de únicamente imperfecto; *ser, tener* "be, have" en la lista del pretérito y del imperfecto) con ejemplos paralelos auténticos.

Con respecto a las fuentes del texto, debido a que varios tipos de literatura y temas están asociados con diversas selecciones de un aspecto verbal, al elegir materiales para los estudiantes, los profesores deben considerar las características y la tendencia de uso dentro de los géneros y temas en los materiales elegidos. Los verbos del

español en tiempo pasado en los cuentos de hadas ofrecen una plataforma adecuada para comparar dos aspectos gramaticales diferentes porque ambas formas aparecen en una base cuasi frecuente (60 % en forma pretérita frente a 40 % en forma imperfecta). Además, los documentos de la ONU que ofrecen tanto textos escritos como orales permiten a los alumnos comparar y contrastar las similitudes y diferencias entre los dos formatos. Por último, aunque el lenguaje utilizado en los textos bíblicos se considera antiguo y el vocabulario y la gramática pueden no ser óptimos para facilitar el aprendizaje de los estudiantes, los temas religiosos pueden atraer la atención de aquellos que estén interesados en la religión.

5. Conclusión

En resumen, en cuanto a la creación del CPEIC, hemos completado la primera etapa en la construcción de un corpus trilingüe. La cantidad total de datos es de aproximadamente 4 millones de palabras. Dicho corpus difiere de los corpus paralelos existentes y se caracteriza por la información etiquetada en POS y la alineación de palabras de los tres idiomas más hablados en todo el mundo: español, inglés y chino. El corpus proporciona una plataforma para estudios académicos y se puede utilizar como una herramienta para ayudar a los instructores en la enseñanza y para facilitar a los estudiantes multilingües el proceso de aprendizaje en la conexión de forma y significado.

Mediante el análisis contrastivo y la traducción del tiempo pasado, el CPEIC obtuvo de manera sistemática y eficiente patrones, rasgos y categorías clasificadas de dos aspectos gramaticales del español, el pretérito y el imperfecto, según la frecuencia de uso. Con base en el resultado del análisis, que incluye evidencia tanto respaldada por estudios anteriores como contradictoria a ellos, la secuencia de aprendizaje y las sugerencias para la selección de textos proporcionan referencias para los maestros y estudiantes de español L3 en Taiwán.

Reconocimiento

Extendemos nuestro agradecimiento al Ministerio de Ciencia y Tecnología de Taiwán por su generoso apoyo a la subvención de este proyecto (MOST 101-2410-H-006-088-MY2), el apoyo técnico del equipo de Ciencias de la Computación e Ingeniería de la Información (CSIE) en la Universidad Nacional Cheng Kung en Taiwán, y a los asistentes de investigación.

Referencias

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies* 7(2). 223–243. 10.1075/target.7.2.03bak.

Bardovi-Harlig, Kathleen. 2000. Tense and aspect in second language acquisition: Form, meaning, and use. *Language Learning: A Journal of Research in Language Studies* 50. 1.

Castillo Rodríguez, Cristina. 2009. La elaboración de un corpus ad hoc paralelo multilingüe. *Tradumàtica* (7). on–line.

Corpas Pastor, Gloria. 2003. TURICOR: compilación de un corpus de contratos turísticos (alemán, español, inglés, italiano) para la generación textual multilingüe y la traducción jurídica. En *Panorama actual de la investigación en traducción e interpretación*, 373–384.

Dimitrova, Ludmila, Violetta Koseska-Toszewa, Danuta Roszko & Roman Roszko. 2010. Application of multilingual corpus in contrastive studies (on the example of the Bulgarian-Polish-Lithuanian parallel corpus). *Cognitive Studies* (10). 10.11649/cs.2010.013.

ELRA, European Language Resources Association. 1996. Multilingual and Parallel Corpora. https://tinyurl.com/ycz4x9ky.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. En *10th Machine Translation Summit*, 79–86.

Lee, David. 2010. Bookmarks for corpus-based linguists. Available from World Wide Web: http://devoted.to/corpora.

Lixun, Wang. 2001. Exploring parallel concordancing in English and Chinese. *Language Learning & Technology* 5(3). 174–184.

Lu, Hui-Chuan, An Chung Cheng & Sheng-Yun Hung. 2015. La adquisición del tiempo-aspecto en L3 para los aprendices taiwaneses. *Círculo de Lingüística Aplicada a la Comunicación* 63. 200–217. 10.5209/rev_CLAC.2015.v63.50175.

Ma, Qing, Kyoko Kanzaki, Yujie Zhang, Masaki Murata & Hitoshi Isahara. 2004. Self-organizing semantic maps and its application to word alignment in Japanese–Chinese parallel corpora. *Neural networks* 17(8-9). 1241–1253. 10.1016/j.neunet.2004.07.011.

Ma, Wei-Yun & Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation

system for the first international Chinese word segmentation bakeoff. En *Second SIGHAN workshop on Chinese language processing*, 168–171. 10.3115/1119250.1119276.

Malmkjær, Kirsten. 2005. *Linguistics and the language of translation*. Edinburgh university press.

McEnery, Tony & Richard Xiao. 2005. The babel English-Chinese parallel corpus. http://www.lancs.ac.uk/fass/projects/corpus/babel/babel.html.

Och, Franz Josef & Hermann Ney. 2000. Improved statistical alignment models. En *38th Annual Meeting on Association for Computational Linguistics*, 440–447. 10.3115/1075218.1075274.

Rabadán, Rosa, Belén Labrador & Noelia Ramón. 2009. Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment English → Spanish. *Babel* 55(4). 303–328. 10.1075/babel.55.4.01rab.

Rabadán, Rosa. 2002. Análisis contrastivo y traducción inglés-español: el programa ACTRES, 35–55.

Robison, Richard E. 1990. The primacy of aspect: Aspectual marking in English interlanguage. *Studies in second language acquisition* 12(3). 315–330. 10.1017/S0272263100009190.

Salaberry, Maximo Rafael. 2002. Tense and aspect in the selection of Spanish past tense verbal morphology. *Language acquisition and language disorders* 27. 397–416. 10.1075/lald.27.16sal.

Salaberry, Maximo Rafael. 2011. Assessing the effect of lexical aspect and grounding on the acquisition of L2 Spanish past tense morphology among L1 English speakers. *Bilingualism: Language and Cognition* 14(2). 184–202. 10.1017/S1366728910000052.

Solorio, Thamar & Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. En *Conference on Empirical Methods in Natural Language Processing*, 1051–1060.

Sun, Le, Song Xue, Weimin Qu, Xiaofeng Wang & Yufang Sun. 2002. Constructing of a large-scale Chinese-English parallel corpus. En *3rd workshop on Asian language resources and international standardization*, 1–8.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. En *Language Resources Evaluation Conference (LREC)*, 2214–2218.

Vendler, Z. 1967. Verbs and times text. *Linguistics in Philosophy. Ithaca, NY: Cornell University Press* 97.

Xiao, Richard & Tony McEnery. 2004. *Aspect in Mandarin Chinese: A corpus-based study*, vol. 73. John Benjamins Publishing.

http://www.linguamatica.com/

linguaMATICA

Artigos de Investigação